Accurate Approximations for Posterior Moments and Marginals

by

Luke Tierney[1]
and
Joseph B. Kadane[2]

Technical Report No. 431

1. School of Statistics, University of Minnesota.
2. Departments of Statistics and Social Sciences, Cargegie-Mellon University. Research sponsored in part by the Office of Naval Research, Contract N00014-82-K-0622.

Abstract

This paper describes approximations to the posterior means and variances of positive functions of a real or vector valued parameter. These approximations can be applied directly to compute approximations to the predictive density, and they can be modified for use in approximating marginal posterior densities in multi-parameter problems. To apply the proposed method one only needs to be able to maximize slightly modified likelihood functions and to evaluate the observed information at the maxima. Nevertheless, the resulting approximations are generally as accurate and in some cases more accurate than more conventional approximations based on third order expansions of the likelihood and requiring either the evaluation of third derivatives or the use of derivative-free maximization prodecures. When used to obtain marginal posterior densities, this method behaves very much like the saddle point approximation method for sampling distributions. In particular, for several distributions, including the normal-gamma distribution and the Dirichlet distribution, the approximations to the marginal densities (renormalized to integrate exactly to one) are exact.

Key words: Bayesian inference, Laplace method.

Abstract

This paper describes approximations to the posterior means and variances of positive functions of a real or vector valued parameter. These approximations can be applied directly to compute approximations to the predictive density, and they can be modified for use in approximating marginal posterior densities in multi-parameter problems. To apply the proposed method one only needs to be able to maximize slightly modified likelihood functions and to evaluate the observed information at the maximum. Nevertheless, the resulting approximations are generally as accurate and in some cases more accurate than more conventional approximations based on third-order expansions of the likelihood and requiring either the evaluation of third derivatives or the use of derivative-free maximization procedures. When used to obtain marginal posterior densities, this method behaves very much like the saddle point approximation method for sampling distributions. In particular, for several distributions including the normal-gamma distribution and the Inverted Wishart distribution, the approximations to the marginal densities (renormalized to integrate exactly to one) are exact.

Key words: Bayesian inference, Laplace method.

1. Introduction and Summary

A user of Bayesian methods in practice needs to be able to evaluate various characteristics of posterior and predictive distributions, especially their densities, means and variances. Unless the problem involves a conjugate prior-likelihood pair, these tasks can not be performed in closed form; analytical or numerical approximation methods are needed. The simplest analytical approximations are based on the asymptotic normality of the posterior distribution. However, this asymptotic distribution does not depend on the prior information, a feature that is particularly undesirable for those interested in assessing the effect on inferences of various different prior distributions. More refined approximations are needed.

Lindley (1980) has proposed some approximations for moments that capture the first order effect of the prior distribution on the posterior mean. This is generally accurate enough, but, as Lindley points out, one of the drawbacks of these approximations is that they require the evaluation of third derivatives. This is often a very tedious task, in particular in p-dimensional problems where the number of partial derivatives that are required is $p(p+1)(p+2)/6$. Mosteller and Wallace (1964, Section 4.6C) suggest a similar approach but introduce a transformation of the parameters to avoid the need for the direct use of third derivatives. However, the proposed transformation depends on the second derivatives of the log-likelihood. A numerical maximization routine for locating the posterior mode of the transformed parameters will therefore require third derivatives

1. Introduction and Summary

A number of Bayesian methods in practice would be able to evaluate various characteristics of posterior and predictive distributions, especially their moments, means and variances. Unless the problem involves a conjugate prior-likelihood pair, these tasks can not be performed in closed form; analytical or numerical approximation methods are needed. The analytical approximations are based on the asymptotic normality of the posterior distribution. However, this asymptotic distribution does not depend on the prior information, a feature that is particularly undesirable for those interested in assessing the effect on inferences of various different prior distributions. More refined approximations are needed.

Lindley (1980) has proposed some approximations for moments that capture the first order effect of the prior distribution on the posterior mean. This is generally accurate enough, but, as Lindley points out, one of the drawbacks of these approximations is that they require the evaluation of third derivatives. This is often a very tedious task, in particular in multidimensional problems where the number of partial derivatives that are required is $p(p+1)(p+2)/6$. Moskaljev and Wallace (1984, Section 4.3) suggest a similar approach but introduce a transformation of the parameters to avoid the need for the direct use of third derivatives. However, the proposed transformation depends on the second derivatives of the log-likelihood. A numerical maximization routine for locating the posterior mode of the transformed parameters will likewise require third derivatives

of the log-likelihood unless a more complicated derivative-free algorighm is used.

Numerical approximation methods provide another alternative to analytic approaches. Two approaches that have received considerable attention are Gauss-Hermite quadrature, used by Naylor and Smith (1982,1983), and Monte Carlo integration using importance sampling, e.g. Zellner and Rossi (1982) and Kloek and Van Dijk (1978). The Gauss-Hermite integration approach is quite inexpensive to use if the number of parameters in the problem is on the order of two or three. For four or more parameters, however, the computing time required will often be large enough to make this approach rather unattractive for use in an interactive data analytic framework. Importance sampling has the advantage that its computing requirements do not increase with the dimensionality of the problem. However, to obtain reliable results one often needs a Monte Carlo sample size of around 10,000 replications, and for problems with nontrivial likelihoods this can lead to computing requirements that again make this approach unattractive for use in an interactive framework. Furthermore, some care is needed in the choice of the importance weight function; an incorrect choice can lead to infinite variances.

In this paper we introduce a new analytical approximation for posterior means and variances of nonnegative parameters, or, more generally, of nonnegative functions of parameters. By a simple modification, discussed at the end of Section 3, our approximation can be adapted for use with parameters taking on

both positive and negative values.

To use the approximation, we only need to be able to evaluate first and second derivatives and maximize slightly modified likelihood functions. For a postive function g, the posterior mean of $g(\theta)$ can be written as

$$(1.1) \qquad E_n[g] = E[g(\theta)|X^{(n)}] = \frac{\int g(\theta) e^{\ell(\theta)} \pi(\theta) d\theta}{\int e^{\ell(\theta)} \pi(\theta) d\theta} ,$$

where ℓ is the log-likelihood function and π is the prior density. An approximation to the denominator integral in (1.1) can be obtained as follows. Let $L(\theta) = (\ell(\theta) + \log\pi(\theta))/n$. If L is essentially unimodal, as is generally the case for moderate and large samples, then by expanding L around its maximum, the posterior mode $\hat{\theta}$, we can approximate L by $L(\hat{\theta}) - (\theta - \hat{\theta})^2/(2\sigma^2)$, where σ^2 is minus the inverse of the second derivative of L at $\hat{\theta}$. Using this approximation for the integrand, we can approximate the integral by

$$\int e^{\ell(\theta)} \pi(\theta) d\theta = \int e^{nL(\theta)} d\theta$$

$$\approx \int e^{nL(\hat{\theta}) - n(\theta-\hat{\theta})^2/(2\sigma^2)} d\theta = \sqrt{2\pi}\, \sigma n^{-1/2} e^{nL(\hat{\theta})} .$$

This approximation is quite standard, and it is used, for example, by Lindley and by Mosteller and Wallace in deriving their results. It can be viewed as an application of the Laplace method for integrals, as described in De Bruijen (1961). The new

posterior expectation and maximize it by the Newton (1964) type method

To make this approximation we could use the delta method of Rao (1973, p. 388) and approximate the posterior expectation of ψ by the expectation of the Taylor series expansion of ψ about $\hat{\theta}$, the posterior mode. If one does this, one obtains the following approximation for a posterior expectation of a function ψ of the parameter θ based on the likelihood function $L(\theta)$ and prior density $\pi(\theta)$:

$$E^{*}[\psi(\theta)] = E[\psi(\theta)|x_{(n)}] \doteq \frac{\int \psi_{g(\theta)}\pi(\theta)L(\theta)d\theta}{\int \psi_{g(\theta)}\pi(\theta)L(\theta)d\theta} \quad , \qquad (1.1)$$

where $g(\theta)$ can be written as

$$g(\theta) = \log L(\theta) + \log \pi(\theta) \, .$$

Here $L(\theta)$ is the likelihood function and π is the prior density. An approximation to the denominator integral in (1.1) can be obtained as follows. Let $L(\theta) = \exp(g(\theta))$. If L is sufficiently well behaved, as it is for most statistical models, it will have a unique maximum at $\hat{\theta}$ and g will be well approximated near $\hat{\theta}$ by a quadratic function of θ about $\hat{\theta}$. Thus $L(\theta)$ is approximately proportional to $\exp\{-\frac{1}{2}(\theta-\hat{\theta})'\Sigma^{-1}(\theta-\hat{\theta})\}$ where Σ is the inverse of the negative of the second derivative of L at $\hat{\theta}$, $\Sigma = (-\ddot{g}_{\hat{\theta}})^{-1}$. An approximation for the integral $\int \psi_{g(\theta)}\pi(\theta)L(\theta)d\theta$ is then obtained by expanding $\psi_{g(\theta)}\pi(\theta)$ in a Taylor series about $\hat{\theta}$ and integrating term by term. The result is

$$\int \psi_{g(\theta)}\pi(\theta)L(\theta)d\theta = \int \psi_{g(\theta)}\pi(\theta)\exp\{g(\hat{\theta})-\tfrac{1}{2}(\theta-\hat{\theta})'\Sigma^{-1}(\theta-\hat{\theta})\}d\theta$$

$$\doteq (2\pi)^{p/2}|\Sigma|^{1/2}\exp\{g(\hat{\theta})\}\left[\psi_{g(\hat{\theta})}\pi(\hat{\theta}) + \tfrac{1}{2}\operatorname{tr}\ddot{\psi}_{g(\hat{\theta})}\Sigma\pi(\hat{\theta}) + \dot{\psi}_{g(\hat{\theta})}\dot{\pi}(\hat{\theta})\Sigma\right] \, ,$$

where p is the dimension of θ, $\dot{\psi}$ and $\ddot{\psi}$ are the first and second derivatives of ψ with respect to θ, and $\dot{\pi}$ is the first derivative of π. Applying this approximation to both the numerator and denominator integrals in (1.1) and simplifying, one obtains

$$E^{*}[\psi(\theta)] \doteq \psi_{g(\hat{\theta})} + \tfrac{1}{2}\operatorname{tr}\ddot{\psi}_{g(\hat{\theta})}\Sigma + \dot{\psi}_{g(\hat{\theta})}\dot{\pi}(\hat{\theta})\Sigma/\pi(\hat{\theta}) \, .$$

This approximation is derived in detail in Tierney and Kadane (1985). Note that the first term is the posterior mode of ψ, the second term is a correction for the curvature of the posterior density, and the third term is a correction for the skewness of the posterior density.

feature in the approximation proposed in the present paper is in its approach to the numerator integral in (1.1). Instead of expanding the integrand of this integral about the posterior mode $\hat{\theta}$ as well, which is the approach taken by Lindley, we set $L^* = (\log g + \ell + \log \pi)/n$ and apply Laplace's method to the numerator integral $\int \exp\{nL^*(\theta)\}d\theta$ as well. That is, we approximate this integral by

$$\int g(\theta)e^{\ell(\theta)}\pi(\theta)d\theta = \int e^{nL^*(\theta)}d\theta \approx \sqrt{2\pi}\,\sigma^* n^{-1/2} e^{nL^*(\hat{\theta}^*)} ,$$

where $\hat{\theta}^*$ maximizes L^* and σ^{*2} is minus the inverse of the second derivative of L^* at $\hat{\theta}^*$. Thus we obtain the approximation

$$\hat{E}_n[g] \approx \frac{\sigma^*}{\sigma} \exp\{n(L^*(\hat{\theta}^*) - L(\hat{\theta}))\} \tag{1.2}$$

for the posterior mean of $g(\theta)$.

Approximation (1.2) is extremely accurate. Basically, if g is smooth and bounded away from zero near the mode of L, then the functions L and L^* are very similar in shape, and the use of the same approximation method in the numerator and the denominator causes a cancellation of first order error terms. As a result, the relative error of this approximation is of order $O(n^{-2})$. This is the same order of error achieved by Lindley's approximations and the approximations proposed by Mosteller and Wallace, which suffer the additional expense of requiring either the evaluation of third derivatives, or the use of derivative-free optimization methods. Furthermore, if we use (1.2) to approximate the

The approximation proposed in the present paper is an alternative approach to the numerator integral in (1.1). Instead of approximating the integrand of this integral about the posterior mode $\hat{\theta}$ as well, which is the approach taken by Lindley, we now approximate L^* $(= \log p + \log \pi)$ and apply Laplace's method to the numerator integral $\int \exp(nL^*(\theta))d\theta$ as well. That is, we approximate this integral by

$$\int g(\theta)e^{nL^*(\theta)}\pi(\theta)d\theta = \int e^{nL^*(\theta)}d\theta \doteq (2\pi)^{p/2}\sigma_n^{*p}|\Sigma|^{1/2}e^{nL^*(\hat{\theta}^*)},$$

where $\hat{\theta}^*$ maximizes L^* and σ_n^{*2} is minus the inverse of the second derivative of L^* at $\hat{\theta}^*$. Thus we obtain the approximation

$$E(g) \doteq \frac{\hat{\sigma}_n^*}{\hat{\sigma}_n}\exp(n(L^*(\hat{\theta}^*) - L(\hat{\theta}))) \qquad (1.2)$$

for the posterior mean of $g(\theta)$.

The approximation (1.2) is extremely accurate. Basically, if g is smooth and bounded away from zero near the mode of L, then the functions L and L^* are very similar in shape, and the use of the same approximation method in the numerator and the denominator causes a cancellation of first order error terms. As a result, the relative error of this approximation is of order $O(n^{-2})$. This is the same order of error achieved by Lindley's approximations and the approximations proposed by Mosteller and Wallace, which suffer the additional expense of requiring either the evaluation of third derivatives, or the use of derivative-free optimization methods. Furthermore, if we use (1.2) to approximate the

posterior variance $V_n[g]$ of $g(\theta)$ by

$$\hat{V}_n[g] = \hat{E}_n[g^2] - \hat{E}_n[g]^2 \quad , \tag{1.3}$$

then a further cancellation of errors occurs and the resulting relative error is again of order $O(n^{-2})$. By contrast, if $\hat{E}_n[g^2]$ and $\hat{E}[g]$ are replaced by Lindley's or Mosteller and Wallace's approximations then the resulting approximate variance is only accurate up to a relative error of order $O(n^{-1})$.

In multiparameter problems the Laplace method can also be used to approximately integrate out a subset of the parameters to obtain approximate marginal posterior distributions. This was first pointed out by Leonard (1982) in his comment on the paper by Faulkenberry and Lejeune (1982). The behaviour of the resulting approximation is very similar to the behaviour of the saddle point approximation of Daniels (1954) to the sampling distribution of the sample mean. In particular, for several problems, including the normal-gamma distribution and the Dirichlet distribution, the approximation (renormalized to integrate exactly to one) is exact. For other problems, the relative error is generally uniformly of order $O(n^{-1})$ on bounded subintervals of the parameter space, and, in some cases, on the entire parameter space. Furthermore, on $n^{-1/2}$-neighborhoods of the "true" parameter value, that is neighborhoods that shrink at the rate $n^{-1/2}$, the error of the renormalized approximation is of order $O(n^{-3/2})$. By contrast, Edgeworth-type expansions are generally only accurate in

posterior variance $V_n[g]$ of $g(\theta)$ by

$$V_n[g] = E_n[g^2] - E_n[g]^2 \quad . \tag{1.3}$$

then a further cancellation of errors occurs and the resulting relative error is again of order $O(n^{-2})$. By contrast, if $E_n[g^2]$ and $E_n[g]$ are replaced by Lindley's or Mosteller and Wallace's approximations then the resulting approximate variance is only accurate up to a relative error of order $O(n^{-1})$.

In multiparameter problems the Laplace method can also be used to approximately integrate out a subset of the parameters to obtain approximate marginal posterior distributions. This was first pointed out by Leonard (1982) in his comment on the paper by Tierney and Kadane (1982). The behaviour of the resulting approximation is very similar to the behaviour of the saddle point approximation of Daniels (1954) to the sampling distribution of the sample mean. In particular, for several problems, including the normal-gamma distribution and the Dirichlet distribution, the approximation (renormalized to integrate exactly to one) is exact. For other problems, the relative error is generally uniformly of order $O(n^{-1})$ on bounded subintervals of the parameter space, and, in some cases, on the entire parameter space. Furthermore, on "$n^{-1/2}$-neighborhoods of the "true" parameter value, that is neighborhoods that shrink at the rate $n^{-1/2}$, the error of the renormalized approximation is of order $O(n^{-3/2})$. By contrast, Edgeworth-type expansions are generally only accurate in

$n^{-1/2}$-neighborhoods of the true parameter (on fixed length intervals their maximal relative error tends to infinity), and to achieve a local error of order $O(n^{-3/2})$ two terms in addition to the first order normal density are needed.

In the next section we review the Laplace approximation method and obtain expressions for the resulting relative errors. Section 3 presents our approximations for the posterior mean and variance of a nonnegative function of a one-dimensional and a p-dimensional parameter. Section 4 discusses the approximation of predictive distributions and Section 5 presents our results for the approximation of marginal densities. In Section 6 we illustrate our results with two examples. In the first example we compare exact and approximate posterior moments for Poisson data with a gamma prior distribution on the mean. In the second example, we apply our approximation technique to the three-parameter Pareto model of Turnbull, Brown and Hu (1974) for the Stanford heart transplant data. The posterior distribution of this problem has been studied by Naylor and Smith (1982) using Gauss-Hermite quadrature. Section 7 gives some concluding remarks.

In our derivation of relative errors given in Sections 2 through 5 we have chosen to sacrifice rigor for clarity of exposition. Appendix 2 contains rigorous statements and proofs of some of our results.

$n^{-1/2}$-neighborhoods of the true parameter (on fixed width intervals their maximal relative error tends to infinity), and to achieve a local error of order $O(n^{-3/2})$ two terms in addition to the first order normal density are needed.

In the next section we review the Laplace approximation method and obtain expressions for the resulting relative errors. Section 3 presents our approximations for the posterior mean and variance of a monotone function of a one-dimensional and a p-dimensional parameter. Section 4 discusses the approximation of predictive distributions and Section 5 presents our results for the approximation of marginal densities. In Section 6 we illustrate our results with two examples. In the first example we have an exact and approximate posterior moments for Poisson data with a gamma prior distribution on the mean. In the second example we apply our approximation technique to the three-parameter Pareto model of Tarone[1], Brown and Hu (1977A) for the stock-yard bovine tuberculosis data. The population distribution of this problem has been studied by Hayter and Smith (1982) using Pearson-Hermite curves. Section 7 gives some concluding remarks.

In the derivation of relative errors given in Sections 2 through 5 we have chosen to sacrifice rigor for clarity of exposition. Appendix 2 contains rigorous statements and proofs of some of our results.

2. Laplace's Method for Integrals and Ratios of Integrals

Laplace's method for integrals, as described, for example, in DeBruijen (1961), provides an approximation for integrals of the form $\int e^{nL(\theta)}d\theta$ when n is large. The idea is that if L has a unique maximum at $\hat{\theta}$, then for large n the value of this integral depends only on the behaviour of the function L near its maximum. Thus if we set $\sigma^2 = -1/L''(\hat{\theta})$, then we can replace $L(\theta)$ by $L(\hat{\theta}) - (\theta-\hat{\theta})^2/(2\sigma^2)$. This produces the approximation

$$\int e^{nL(\theta)}d\theta \approx e^{nL(\hat{\theta})}\int \exp\{-\frac{n(\theta-\hat{\theta})^2}{2\sigma^2}\}d\theta$$

$$= \sqrt{2\pi}\,\sigma\, n^{-1/2}\exp\{nL(\hat{\theta})\}\ .$$

By expanding $n(L(\theta) - L(\hat{\theta}) + (\theta - \hat{\theta})^2/(2\sigma^2)$ about $\hat{\theta}$ and e^x about zero, it is possible to obtain the more refined result

$$(2.1) \qquad \int e^{nL(\theta)}d\theta = \sqrt{2\pi}\,\sigma\, n^{-1/2}e^{nL(\hat{\theta})}(1 + \frac{a}{n} + \frac{b}{n^2} + O(n^{-3}))\ ,$$

where, setting $f^{(k)}(x) = (\frac{d}{dx})^k f(x)$ and $\mu_k = \frac{k!}{(k/2)!2^{k/2}}$

for even k, the constansts a and b are given by

$$a = \mu_4 \frac{\sigma^4}{24} L^{(4)}(\hat{\theta}) + \mu_6 \frac{\sigma^6}{72} L^{(3)}(\hat{\theta})^2$$

and

$$b = \mu_6 \frac{\sigma^6}{720} L^{(6)}(\hat{\theta})^2 + \mu_8 \frac{\sigma^8}{1152} L^{(4)}(\hat{\theta})^2$$
$$+ \mu_{10} \frac{\sigma^{10}}{1728} L^{(3)}(\hat{\theta})^2 L^{(4)}(\hat{\theta})$$
$$+ \mu_{12} \frac{\sigma^{12}}{31104} L^{(3)}(\hat{\theta})^4.$$

Result (2.1) remains valid if L is replaced by a sufficently well behaved sequence L_n of functions. In this case the coefficients a and b may depend on n, but this dependence will be suppressed. If a and b do indeed depend on n, we will assume regularity conditions for the sequence L_n that insure that a and b are bounded in n. Theorem 1 of Appendix 2 formalizes this extension.

As an example, consider the function $L(x) = \log(x) - x/n$. The maximum of this function occurs at $\hat{x} = n$. So $\sigma = \sqrt{-1/L''(\hat{x})} = \hat{x} = n$, and Laplace's method applied to the integral $\int_0^\infty e^{nL(x)}dx$ yields

$$n! = \int_0^\infty x^n e^{-x} dx = \int_0^\infty e^{nL(x)} dx \approx \sqrt{2\pi}\, n^{n+1/2} e^{-n},$$

which is Stirling's approximation. Equation (2.1) shows that

$$(2.2) \qquad n! = \sqrt{2\pi}\, n^{n+1/2} e^{-n}\left(1 + \frac{1}{12n} + \frac{1}{288n^2} + O(n^{-3})\right),$$

which is a well-known expression for the error of Stirling's approximation (see, for example, the remarks following equation (7.6.27) in Wilks (1962)).

These results can be generalized to multiple integrals as

well; for simplicity we only give the first order error term in this case: If L is a function from $\mathbb{R}^p$ to $\mathbb{R}$ with its maximum at $\hat{\theta}$, then

$$(2.3) \qquad \int e^{nL(\theta)}d\theta = (2\pi/n)^{p/2}(\det \Sigma)^{1/2}e^{nL(\hat{\theta})}\ (1 + \frac{a}{n} + O(n^{-2})),$$

where Σ is minus the inverse of the Hessian of L at $\hat{\theta}$. Setting $L_i = \frac{\partial}{\partial\theta_i} L(\hat{\theta})$, $L_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j} L(\hat{\theta})$, etc. and denoting the elements of Σ by σ_{ij}, the constant a for (2.3) is given by

$$\begin{aligned} a = {} & \frac{1}{24} \sum_{ijkl} (\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})L_{ijkl} \\ & + \frac{1}{72} \sum_{ijklmr} (\sigma_{ij}\sigma_{kl}\sigma_{mr} + \sigma_{ij}\sigma_{km}\sigma_{lr} + \sigma_{ij}\sigma_{kr}\sigma_{lm} \\ & + \sigma_{ik}\sigma_{jl}\sigma_{mr} + \sigma_{ik}\sigma_{jm}\sigma_{lr} + \sigma_{ik}\sigma_{jr}\sigma_{lm} \\ & + \sigma_{il}\sigma_{jk}\sigma_{mr} + \sigma_{il}\sigma_{jm}\sigma_{kr} + \sigma_{il}\sigma_{jr}\sigma_{km} \\ & + \sigma_{im}\sigma_{jk}\sigma_{lr} + \sigma_{im}\sigma_{jl}\sigma_{kr} + \sigma_{im}\sigma_{jr}\sigma_{kl} \\ & + \sigma_{ir}\sigma_{jk}\sigma_{lm} + \sigma_{ir}\sigma_{jl}\sigma_{km} + \sigma_{ir}\sigma_{jm}\sigma_{kl})L_{ijk}L_{lmr}. \end{aligned}$$

A more detailed derivation of this constant is given in Appendix 1.

A posterior expectation $E_n[g]$ of a positive function g is a ratio of two integrals of the form $\int\exp\{nL^*(\theta)\}d\theta / \int\exp\{nL(\theta)\}d\theta$ where the difference $L^* - L$ is of order $O(n^{-1})$. Due to this small difference the two integrands are very similar in shape. Thus, if Laplace's method is applied to both numerator and denominator integrals, then the first order approximation errors cancel, resulting in an error of order

convergence assumptions to an order of magnitude

asymptotic independence, then the first order approximation errors

[illegible] is established, a method is sketched for both unbiased and

one) difference form two independent such variables in which

terms for two independent of the form $\int_{\alpha}^{\alpha} \text{tr}(\Omega(\theta)) d\theta$

a bootstrap expectation $E^{*}[\cdot]$ of a bootstrap function ϕ is a

where Ω.

a more detailed derivation of this coefficient is given in

$$+ \alpha^{jl}\alpha^{ik}\alpha^{jm} + \alpha^{jk}\alpha^{il}\alpha^{jm} + \alpha^{jk}\alpha^{jl}\alpha^{im})\Gamma^{ijk}\Gamma^{lmn},$$

$$+ \alpha^{jm}\alpha^{il}\alpha^{kn} + \alpha^{jm}\alpha^{ik}\alpha^{ln} + \alpha^{jm}\alpha^{jk}\alpha^{in}$$

$$+ \alpha^{jl}\alpha^{ik}\alpha^{mn} + \alpha^{jl}\alpha^{im}\alpha^{kn} + \alpha^{jl}\alpha^{jk}\alpha^{in}$$

$$+ \alpha^{jk}\alpha^{il}\alpha^{mn} + \alpha^{jk}\alpha^{im}\alpha^{ln} + \alpha^{jk}\alpha^{jl}\alpha^{in}$$

$$+ \frac{1}{3!}\sum_{i,j,k,l,m,n} (\alpha^{il}\alpha^{jk}\alpha^{mn} + \alpha^{il}\alpha^{jm}\alpha^{kn} + \alpha^{il}\alpha^{jn}\alpha^{km})$$

$$= \frac{n}{4}\sum_{i,j,k,l} (\alpha^{ij}\alpha^{kl} + \alpha^{ik}\alpha^{jl} + \alpha^{il}\alpha^{jk})\Gamma^{ijkl}$$

this element of $\mathcal{I}$ by α^{ij}, the coefficient of (α^{ij}) is given by

Define $L_i = \frac{\partial}{\partial\theta^i} L(\theta)$, $L_{ij} = \frac{\partial}{\partial\theta^i\partial\theta^j} L(\theta)$, etc., and define

where $\mathcal{I}$ is minus the inverse of the Hessian of L at $\hat{\theta}$.

$$\int s_{ul}(\theta)\, d\theta = (2\pi/n)^{p/2} |\det \mathcal{I}|^{1/2} g_{ul}(\theta) \quad (1 = \frac{u}{n} + O(n^{-2})), \tag{2.2}$$

by [illegible]

[illegible] is a function (from $\mathbb{R}_p$ to $\mathbb{R}$ with) [illegible] maximum at

$O(n^{-2})$ for the ratio approximation. That is, if L satisfies (2.1) and $L^* = W/n + L$, then

$$(2.4) \qquad \frac{\int e^{nL^*(\theta)}d\theta}{\int e^{nL(\theta)}d\theta} = \frac{\sigma^* e^{nL^*(\hat{\theta}^*)}}{\sigma e^{nL(\hat{\theta})}}\left[1 + \frac{c}{n^2} + O(n^{-3})\right],$$

where $\hat{\theta}^*$ maximizes L^*, $\sigma^{*2} = -1/L^{*\prime\prime}(\hat{\theta}^*)$,

$$c = W'(\hat{\theta})d_1 + W''(\hat{\theta})d_2 + W'''(\hat{\theta})d_3 + W^{iv}(\hat{\theta})d_4 ,$$

and the d_i's are given by

$$d_1 = \frac{1}{24}\mu_4\sigma^6 L_5 + \frac{1}{12}\mu_4\sigma^8 L_3 L_4 + \frac{1}{36}\mu_6\sigma^8 L_3 L_4 + \frac{1}{24}\mu_6\sigma^{10} L_3^3,$$
$$d_2 = \frac{1}{12}\mu_4\sigma^6 L_4 + \frac{1}{24}\mu_6\sigma^8 L_3^2,$$
$$d_3 = \frac{1}{36}\mu_6\sigma^6 L_3$$

and

$$d_4 = \frac{1}{24}\mu_4\sigma^4 .$$

To see this, apply (2.1) to $\int \exp\{nL^*(\theta)\}d\theta$ and $\int \exp\{nL(\theta)\}d\theta$ to obtain

$$\frac{\int e^{nL^*(\theta)}d\theta}{\int e^{nL(\theta)}d\theta} = \frac{\sigma^*}{\sigma}\exp\{n(L^*(\hat{\theta}^*) - L(\hat{\theta}))\}\frac{(1 + \frac{a^*}{n} + \frac{b^*}{n^2} + O(n^{-3}))}{(1 + \frac{a}{n} + \frac{b}{n^2} + O(n^{-3}))}$$

$$= \frac{\sigma^*}{\sigma}\exp\{n(L^*(\hat{\theta}^*) - L(\hat{\theta}))\}(1 + \frac{a^*-a}{n} + \frac{b^*-b-a(a^*-a)}{n^2} + O(n^{-3})),$$

again suppressing any dependence of a^*, b^* and a, b on n.

Then observe that $\hat{\theta}^*$ solves $L^{*\prime}(\theta) = 0$ and $\hat{\theta}$ solves $L'(\theta) = 0$, and thus

$$\begin{aligned} 0 = L^{*\prime}(\hat{\theta}^*) &= L'(\hat{\theta}^*) + \frac{1}{n} W'(\hat{\theta}^*) \\ &\approx L'(\hat{\theta}) + (\hat{\theta}^* - \hat{\theta}) L''(\hat{\theta}) + \frac{1}{n} W'(\hat{\theta}) \\ &= -(\hat{\theta}^* - \hat{\theta})/\sigma^2 + \frac{1}{n} W'(\hat{\theta}). \end{aligned}$$

So $\hat{\theta}^* - \hat{\theta} = \frac{1}{n} W'(\hat{\theta})\sigma^2 + O(n^{-2})$. Together with the fact that $L^*(\theta) - L(\theta) = \frac{1}{n} W(\theta) = O(n^{-1})$ for any θ, this implies that $a^* - a$ and $b^* - b$ are both of order $O(n^{-1})$. Further details are given in Theorem 2 of Appendix 2.

The multidimensional analog of (2.4) is

$$(2.5) \qquad \frac{\int e^{nL^*(\theta)} d\theta}{\int e^{nL(\theta)} d\theta} = \left(\frac{\det \Sigma^*}{\det \Sigma}\right)^{1/2} \exp\{n(L^*(\hat{\theta}^*) - L(\hat{\theta}))\} \cdot \left(1 + \frac{c}{n^2} + O(n^{-3})\right).$$

Setting $W_i = \frac{\partial}{\partial \theta_i} W(\hat{\theta})$, $W_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} W(\hat{\theta})$, etc., the coefficient c in (2.5) can be written as

$$c = \sum_i d_i W_i + \sum_{ij} d_{ij} W_{ij} + \sum_{ijk} d_{ijk} W_{ijk} + \sum_{ijkl} d_{ijkl} W_{ijkl},$$

where the coefficients d_i, d_{ij} etc. are independent of W.

To apply (2.4) or (2.5), we need to be able to evaluate and maximize L and L^* and to evaluate the second derivatives of L and L^*. as a practical point, once $\hat{\theta}$, the location of the maximum of L, has been determined, it can be used as a starting value for a numerical search for $\hat{\theta}^*$, the maximum of L^*. Generally, the number of iterations needed to find $\hat{\theta}^*$ from $\hat{\theta}$ will be quite small. In fact, since the asymptotic statements of (2.4) and (2.5) only depend on the fact that $\hat{\theta}^* - \hat{\theta} = \frac{1}{n} W'(\hat{\theta})\sigma^2 + O(n^{-2})$, they remain valid if we replace $\hat{\theta}^*$ by $\hat{\theta}' = \hat{\theta} + \frac{1}{n} + W'(\hat{\theta})\sigma^2$, a single Newton step from $\hat{\theta}$ towards $\hat{\theta}^*$.

3) Posterior Means and Variances

The results of the previous section imply that, under certain mild regularity conditions, the posterior mean $E_n[g]$ of a function $g(\theta)$ that is positive at the true parameter value θ_o can be approximated by $\hat{E}_n[g]$ given in (1.2). By (2.4), the error of this approximation is described by

$$(3.1) \qquad E_n[g] = \hat{E}_n[g]\left(1 + \frac{c}{n^2} + O(n^{-3})\right) ,$$

where

$$(3.2) \qquad c = G_1 d_1 + G_2 d_2 + G_3 d_3 + G_4 d_4 ,$$

$G = \log g$, $G_k = \left(\frac{d}{d\theta}\right)^k G(\hat{\theta})$, and the coefficients $d_1, \ldots, d_4$, are bounded in n (their dependence on n has again been supressed) and independent of G. For a detailed justification, along with a statement of regularity conditions, see Theorem 3 of Appendix 2.

A similar approximation applies in the multiparameter case: Set

$$(3.3) \qquad \hat{E}_n[g] = \left(\frac{\det \Sigma^*}{\det \Sigma}\right)^{1/2} \exp\{n(L^*(\hat{\theta}^*) - L(\hat{\theta}))\} ,$$

where $\hat{\theta}^*$ and $\hat{\theta}$ maximize L^* and L, respectively, and Σ^* and Σ are minus the inverse Hessians of L^* and L at $\hat{\theta}^*$ and $\hat{\theta}$, respectively. By (2.5) the error of this approximation is described by

$$(3.4) \qquad E_n[g] = \hat{E}_n[g]\left(1 + \frac{c}{n^2} + O(n^{-3})\right) ,$$

where

$$c = \sum_i G_i\, d_i + \sum_{ij} G_{ij}\, d_{ij} + \sum_{ijk} G_{ijk}\, d_{ijk} + \sum_{ijkl} G_{ijkl}\, d_{ijkl},$$

$G = \log g$, $G_i = \frac{\sigma}{\sigma\theta_i} G(\hat{\theta})$, $G_{ij} = \frac{\sigma^2}{\sigma\theta_i \sigma\theta_j} G(\hat{\theta})$, etc, and the coefficients d_i, d_{ij}, etc are bounded and independent of G.

Approximations (1.2) or (3.3) can also be used to approximate the posterior variance $V_n[g]$ of g. Applying (1.2) to both $E_n[g^2]$ and $E_n[g]$, we obtain the approximation

(3.5) $$\hat{V}_n[g] = \hat{E}_n[g^2] - \hat{E}_n[g].$$

By (3.1) we might expect the absolute error of this approximation to be $O(n^{-2})$, and since $V_n[g] \sim (nI(\theta_0))^{-1}$, where $I(\theta_0) = E\left[\left(\frac{\partial}{\partial\theta} \log f(x|\theta_0)\right)^2 | \theta = \theta_0\right]$ is the expected Fisher information for the density f at the true parameter θ_0, this would produce a relative error of order $O(n^{-1})$. This is the same order of error as the error of σ^2, the inverse of the observed information at the posterior mode. In fact, however, a phenomenon similar to the one observed when using Laplace's method for ratios of integrals occurs: The first order error terms cancel, and we have

$$V_n[g] = \hat{V}_n[g](1 + O(n^{-2})).$$

To see this, note that $\log g^2 = 2 \log g$; thus if

$$E_n[g] = \hat{E}_n[g](1 + \frac{c}{n^2} + O(n^{-3})) ,$$

then by (3.2)

$$E_n[g^2] = \hat{E}_n[g^2](1 + \frac{2c}{n^2} + O(n^{-3})) .$$

Hence

$$\begin{aligned} V_n[g] &= E_n[g^2] - E_n[g]^2 \\ &= \hat{E}_n[g^2](1 + \frac{2c}{n^2} + O(n^{-3})) - \hat{E}_n[g]^2(1 + \frac{c}{n^2} + O(n^{-3}))^2 \\ &= \hat{E}_n[g^2](1 + \frac{2c}{n^2} + O(n^{-3})) - \hat{E}_n[g]^2(1 + \frac{2c}{n^2} + O(n^{-3})) \\ &= (\hat{E}_n[g^2] - \hat{E}_n[g]^2)(1 + \frac{2c}{n^2}) + O(n^{-3}) \\ &= \hat{V}_n[g](1 + \frac{2c}{n^2}) + O(n^{-3}) . \end{aligned}$$

Since $V_n[g] \sim I(\theta_0)^{-1}/n$, we have $\hat{V}_n[g] \sim I(\theta_0)^{-1}/n$ as well; thus the final $O(n^{-3})$ error term can be written as $\hat{V}_n[g]O(n^{-2})$, and we have

$$V_n[g] = \hat{V}_n[g](1 + O(n^{-2})) ,$$

as claimed.

A similar calculation shows that if we approximate the posterior covariance $C_n[g,h]$ of two positive functions $g(\theta)$ and $h(\theta)$ by

(3.6) $$\hat{C}_n[g,h] = \hat{E}_n[gh] - \hat{E}_n[g]\hat{E}_n[h] \quad ,$$

then

(3.7) $$C_n[g,h] = \hat{C}_n[g,h](1 + O(n^{-2}))$$

as well.

As a practical point, it is worth mentioning that approximation (3.5) should be used with caution if n is very large, since it involves the computation of a small number as the difference between two large numbers. However, if computations are done with sufficient precision, then for most practical sample sizes this will not cause any problems. If the sample sizes really are very large, then $\frac{1}{n}\,\sigma_n^2 g'(\hat{\theta}_n)^2$ will generally be sufficiently accurate as an approximation to $V_n[g]$. At the other end of the spectrum, if n is very small then it is possible for the variance approximation (3.5) to be negative, and for a covariance matrix computed from (3.6) not to be positive semidefinite. This should be checked in any application, but in most cases it does not seem to be a problem even for moderate sample sizes. More work is needed to see if modified variance approximations can be obtained that are guaranteed to be positive.

Lindley (1980) suggests an alternate approximation to $E_n[g]$ that is obtained by taking all terms of order $O(n^{-1})$ or less from an expansion of numerator and denominator integrands in (1.1) about the MLE or the posterior mode. The error in this approximation is of order $O(n^{-2})$, as is the error of the

approximations proposed in the present paper. However, to use it we need to evaluate the third derivatives of the log-likelihood. For a one parameter problem this is generally not serious, but as p increases the number of third derivatives that need to be computed is $p(p+1)(p+2)/6$, which rapidly becomes prohibitive.

Mosteller and Wallace (1964) propose a similar approximation based on an expansion about the posterior mode relative to a suitably chosen function $b(\theta)$. That is, they write the posterior density $\pi_n(\theta) = \pi(\theta|X^{(n)})$ as $\pi_n(\theta) \propto b(\theta)e^{h(\theta)}$ and use the mode of the function h. The function b can be thought of as the determinant of the Jacobian of a transformation of the parameters. If $b(\theta)$ is chosen to be $b(\theta) = (\partial^2 \log \pi_n(\theta)/\partial\theta^2)^{1/2}$, then the corresponding transformed parameters have zero asymptotic skewness and as a result the term in the approximation involving third derivatives of the log-likelihood vanishes. Thus for this choice of the function b the approximation proposed by Mosteller and Wallace only requires the evaluation of second derivatives of the log-likelihood at the mode of h. However, to compute the mode of $h = \log \pi_n - \log b$ by a gradient-based algorithm requires the evaluation of the derivative of b. For the choice of b given above, this in turn requires the evaluation of the third derivatives of the log-likelihood. Thus the only way to aviod computing these third derivatives in obtaining this approximation is to use a derivative-free optimization method to compute the mode of h. Several such algorighms are available, but they are generally considerably more difficult to use than, say, the Newton-Raphson algorithm.

The asymptotic errors of both Lindley's and of Mosteller and Wallace's approximation methods are of order $O(n^{-2})$, as are the errors of (1.2) and (3.3). However, in certain contexts (1.2) and (3.3) are more accurate. For example, if either Lindley's or Mosteller and Wallace's approximations are used in (3.5) in place of (1.2) or (3.3), then the resulting approximate variances that are obtained generally have an absolute error of order $O(n^{-2})$ and a relative error of order $O(n^{-1})$ instead of the relative error of order $O(n^{-2})$ for (3.5) derived above. Furthermore, in all cases that have been tried so far, both of these methods produce less accurate approximate predictive distributions than the method of this paper. Both of these two alternative methods do, however, have one advantage over the method proposed here: They are directly applicable even if the posterior distribution of $g(\theta)$ is not concentrated almost entirely on the positive (or negative) half line.

A comment on the assumed positivity of g is appropriate. This assumption is needed to insure that the numerator and denominator integrands in (1.1) are similar in shape. This similarity in shape, in turn, is responsible for the cancellation of error terms in the approximation to the ratio (1.1). Thus for the approximation to be accurate for functions g taking both positive and negative values, the posterior distribution of g must be concentrated almost entirely to one side of the origin. If this is not the case, then our approach is not directly applicable. However, a simple modification is available: replace g by $\bar{g} = g + c$ for a suitably large constant c,

approximate $E_n[\bar{g}]$ by $\hat{E}_n[\bar{g}]$, and subtract c from this approximation. It should be sufficient to choose c on the order of five to ten asymptotic standard deviations of g. How well this approach works in practice needs to be investigated more carefully. So far, we have only used it to approximate the mean and variance of a standard normal random variable, and in this case it appears to work quite well.

4) Predictive Distributions

An application of approximation (1.2) worth special attention is to approximating the predictive density

$$f_n(z) = f(z|x^{(n)}) = E_n[f(z|\theta)]$$

at specific values of z. Direct application of (1.2) produces the approximation

$$\hat{f}_n(z) = \hat{E}_n[f(z|\theta)] ,$$

which, by (3.1), has an error of order $O(n^{-2})$. This is similar to the approximation considered by Leonard (1982) in his comment on the paper by Lejeune and Faulkenberry (1982) who propose a similar approximation from the frequentist point of view.

With its error of order $O(n^{-2})$, the approximation $\hat{f}_n(z)$ has a lower order error than Dunsmore's (1976) modification of the simple approximation $f(z|\hat{\theta})$, where $\hat{\theta}$ is the MLE. Dunsmore's approximation includes some but not all terms of order $O(n^{-1})$. In particular, he drops the integral of the second term in the integrand of his equation (1); Lindley's (1980) equation (7) shows that this integral is generally of order $O(n^{-1})$ but no smaller.

5) Marginal Posterior Densities

Laplace's method can also be used to approximate marginal posterior densities of individual parameters in multiparameter settings. The resulting approximation provides a useful alternative to generally more time consuming numerical or Monte Carlo integration techniques. This approach appears to have been first suggested by Leonard (1982). In its use of Laplace's method to integrate out one or more variables from a multivariate function to obtain a density, this approach is also somewhat similiar to the saddle point method introduced by Daniels (1954) and studied further, for example, in Barndorff-Nielsen and Cox (1979) and Daniels (1980).

To obtain our approximation, set $\theta = (\theta_1, \ldots, \theta_p) = (\theta_1, \theta_2)$, i.e. partition the p-vector θ into its first component and the (p-1)-vector of the remaining components. Suppose $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ maximizes πe^{ℓ}, i.e. $\hat{\theta}$ is the posterior mode, and let Σ be minus the inverse of the Hessian of $(\ell + \log \pi)/n$ at $\hat{\theta}$; thus Σ is a $p \times p$ matrix. For a given θ_1, let the (p-1) vector $\hat{\theta}_2^* = \hat{\theta}_2^*(\theta_1)$ maximize the function $h(\cdot) = \pi(\theta_1, \cdot)e^{\ell(\theta_1, \cdot)}$, the function πe^{ℓ} with θ_1 held fixed, and let $\Sigma^* = \Sigma^*(\theta_1)$ be minus the inverse of the Hessian of $(\log h(\cdot))/n$, a $(p-1) \times (p-1)$ matrix. Applying Laplace's method to the integrals in the numerator and denominator of the expression

$$\pi_1(\theta_1 | x^{(n)}) = \pi_{n,1}(\theta_1) = \frac{\int \pi(\theta_1, \theta_2) e^{\ell(\theta_1, \theta_2)} \, d\theta_2}{\int \pi(\theta) e^{\ell(\theta)} \, d\theta}$$

3.1. Marginal Posterior Densities

Laplace's method can also be used to approximate marginal posterior densities of individual parameters in multiparameter settings. The resulting approximation provides a useful alternative to generally more time consuming numerical or Monte Carlo integration techniques. This approach appears to have been first suggested by Leonard (1982). In the use of Laplace's method to integrate out one or more variables from a multivariate function to obtain a density, this approach is also somewhat similar to the modified point method introduced by Druzdzel (1989) and studied further by Thorp, for example, in Bernardo-Ibáñez and Cao (1993) and Davila (1990).

To obtain our approximation, set $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) = (\theta_1, \boldsymbol{\theta}_2)$, i.e., partition the p-vector $\boldsymbol{\theta}$ into its first component and the $(p-1)$-vector of the remaining components. Suppose $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\boldsymbol{\theta}}_2)$ maximizes $m\varepsilon_\Sigma$, i.e. $\hat{\boldsymbol{\theta}}$ is the posterior mode, and let $\hat{\Sigma}$ be minus the inverse of the Hessian of $\log m$ at $\hat{\boldsymbol{\theta}}$; thus $\hat{\Sigma}$ is a $p \times p$ matrix. For a given θ_1, let the $(p-1)$ vector $\boldsymbol{\theta}_2^*(\theta_1) = \boldsymbol{\theta}_2^*$ maximize the function $h(\cdot) = m(\theta_1, \cdot)$ i.e. $m\varepsilon_\Sigma(\theta_1, \boldsymbol{\theta}_2^*(\theta_1))$, with θ_1 held fixed, and let $\hat{\Sigma}_2^* = \hat{\Sigma}_2^*(\theta_1)$ be minus the inverse of the Hessian of $(\log h(\cdot))/n$, a $(p-1) \times (p-1)$ matrix. Applying Laplace's method to the integrals in the numerator and denominator of the expression

$$\pi_1(\theta_1 | x^{(n)}) = \pi_{1,1}(\theta_1) = \frac{\int \pi(\theta_1, \boldsymbol{\theta}_2) e^{\ell(\theta_1, \boldsymbol{\theta}_2)} \, d\boldsymbol{\theta}_2}{\int \pi(\boldsymbol{\theta}) e^{\ell(\boldsymbol{\theta})} \, d\boldsymbol{\theta}}$$

for the marginal posterior density of θ_1 we obtain the approximation

$$(5.1) \qquad \hat{\pi}_{n,1}(\theta_1) = \left(\frac{\det \mathfrak{I}^*(\theta_1)}{2\pi n \det \mathfrak{I}}\right)^{1/2} \frac{\pi(\theta_1, \hat{\theta}_2^*) e^{\mathfrak{l}(\theta_1, \hat{\theta}_2^*)}}{\pi(\hat{\theta}) e^{\mathfrak{l}(\hat{\theta})}}$$

There are several ways to characterize the accuracy of approximation (5.1). One is to use (2.1) and (2.3) to obtain

$$(5.2) \qquad \pi_{n,1}(\theta_1) = \hat{\pi}_{n,1}(\theta_1)(1 + O_{\theta_1}(n^{-1})) ,$$

where $O_{\theta_1}(n^{-1})$ is of order $O(n^{-1})$ but depends on θ_1. Under the assumptions of Theorem 3 of Appendix 2, there is some fixed neighborhood of the true parameter value $\theta_{0,1}$, over which the term $O_{\theta_1}(n^{-1})$ in (5.2) is uniformly of order $O(n^{-1})$. By contrast, the absolute error of the marginal density obtained from Walker's (1969) asymptotic normal approximation is only $O(n^{-1/2})$. Moreover, the relative error of that approximation is only $O(n^{-1/2})$ on neighborhoods that shrink at rate $n^{-1/2}$ towards $\hat{\theta}_1$, the first component of the posterior mode. Similarly, a third order Edgeworth-type approximation, such as the one considered in Tierney (1983), obtained by including third order terms of the expansion of the log-likelihood about the MLE or the posterior mode, has an absolute error of order $O(n^{-1})$. But again the relative error is only $O(n^{-1})$ in $n^{-1/2}$-neighborhoods of $\hat{\theta}_1$. On any interval of fixed length the maximal relative error gernerally tends to infinity.

The main reason that the error in (5.2) is as large as $O(n^{-1})$ is that the dimensionalities of the two integrals in the numerator and the denominator of $\pi_{n,1}$ are different. In fact, most of this error is due to the constant of integration; the error in the approximation of the functional form of $\pi_{n,1}(\theta_1)$ is only of order $O(n^{-3/2})$ in $n^{-1/2}$-neighborhoods of $\hat{\theta}_1$. To see this, fix u, let $\theta_1 = \hat{\theta}_1 + n^{-1/2}u$, and consider the error in the ratio $\hat{\pi}_{n,1}(\theta_1)/\hat{\pi}_{n,1}(\hat{\theta}_1)$ as an approximation to $\pi_{n,1}(\theta_1)/\pi_{n,1}(\hat{\theta}_1)$. First, by (2.3) we can write

$$\pi_{n,1}(\hat{\theta}_1) = \hat{\pi}_{n,1}(\hat{\theta}_1)(1 + \frac{a_n}{n} + O(n^{-2}))$$

and

$$\pi_{n,1}(\theta_1) = \hat{\pi}_{n,1}(\theta_1)(1 + \frac{a_n^*}{n} + O(n^{-2})) \ .$$

Since $\theta_1 - \hat{\theta}_1$ is of order $O(n^{-1/2})$, it is easy to see that $\hat{\theta}_2^* - \hat{\theta}_2$ is of order $O(n^{-1/2})$ as well. As a result, arguing as in the derivation of (2.4), we have $a_n^* - a_n = O(n^{-1/2})$, and therefore

$$\frac{\pi_{n,1}(\theta_1)}{\pi_{n,1}(\hat{\theta}_1)} = \frac{\hat{\pi}_{n,1}(\theta_1)}{\hat{\pi}_{n,1}(\hat{\theta}_1)}(1 + \frac{a_n^* - a_n}{n} + O(n^{-2}))$$

$$= \frac{\hat{\pi}_{n,1}(\theta_1)}{\hat{\pi}_{n,1}(\hat{\theta}_1)}(1 + O(n^{-3/2})) \ ,$$

as claimed. Note that even for a third order Edgeworth-type expansion the error of the corresponding ratio is $O(n^{-1})$; to

achieve an error of order $O(n^{-3/2})$ fourth order terms have to be included and fourth derivatives have to be evaluated.

The main implication of this result is that it suggests that an approximate marginal density computed by (5.1) should be renormalized by numerical integration to integrate to one. This again parallels an observation of Daniels (1956) concerning the saddle point approximation.

To appreciate just how accurately (5.1) can capture the functional form of $\pi_{n,1}(\theta_1)$, consider, for example, the normal-gamma conjugate distribution for normal data with unknown mean and precision. Thus the joint posterior for the mean m and the precision r is of the form

$$\pi(m,r) \propto r^{\alpha-1/2} e^{-r(\beta + \tau(m-\mu)^2/2)}$$

for some α, β, μ and τ. Leonard (1982) points out that (5.1) is remarkably accurate in this case. In fact, a simple calculation shows that (5.1) yields

$$\hat{\pi}_1(m) = \left(\frac{\alpha-1/2}{2\alpha\pi}\,\frac{\alpha\tau}{\beta}\right)^{1/2}\left(1 + \frac{\alpha\tau}{\beta}\,\frac{(m-\mu)^2}{2\alpha}\right)^{-(2\alpha-1)/2}$$

$$\propto \left(1 + \frac{\alpha\tau}{\beta}\,\frac{(m-\mu)^2}{2\alpha}\right)^{-(2\alpha-1)/2}$$

and

$$\hat{\pi}_2(r) = \frac{\beta^{\alpha} r^{\alpha-1} e^{-\beta r}}{\sqrt{2\pi}(\alpha-1/2)^{\alpha-1/2} e^{-\alpha+1/2}} \propto r^{\alpha-1} e^{-\beta r} \;;$$

thus in both cases (5.1) produces the correct functional forms

exactly!

Another joint distribution $\pi(\theta_1, \ldots, \theta_p)$ for which (5.1) produces the exact functional forms of the marginals is the Dirichlet disribution. It would be interesting to obtain a characterization of all joint distributions for which this occurs. A similar phenomenon occurs for the saddle point approximation; that is, there are certain distributions for which that approximation produces the exact functional forms. For the saddle point approximation, Daniels (1980) has obtained a characterization of all cases in which the approximation produces exact results; these turn out to be the normal, gamma and inverse normal distributions. It may be possible to modify Daniels' approach to characterize the joint distributions for which (5.1) produces exact functional forms.

6. Applications

In this section we present two applications of our results. First we approximate the posterior mean and variance of a Poisson mean parameter with a gamma prior distribution and compare the results to the exact values. In the second example we compute posterior means, variances and marginal densities of the three parameters of the Pareto model of Turnbull, Brown and Hu (1974) for the Stanford heart transplant data. These approximations are compared to numerical results obtained by Naylor and Smith (1982).

We begin with the Poisson example with a posterior density proportional to $\lambda^{\Sigma X_i+\alpha-1}e^{-(\beta+n)\lambda} = \lambda^{\tilde{\alpha}-1}e^{-\tilde{\beta}\lambda}$ where $\hat{\alpha} = \Sigma X_i + \alpha$ and $\hat{\beta} = \beta+n$. Thus the posterior mean and variance are given by $\frac{\tilde{\alpha}}{\tilde{\beta}}$ and $\frac{\tilde{\alpha}}{\tilde{\beta}^2}$, respectively. For approximating the posterior mean of λ, set $L(x) = (\tilde{\alpha} - 1)\log x - \tilde{\beta}x$, $\hat{x} = \frac{\tilde{\alpha} - 1}{\tilde{\beta}}$ and $\sigma^2 = \frac{\tilde{\alpha} - 1}{\tilde{\beta}^2}$ for the denominator, and $L^*(x) = \tilde{\alpha}\log x - \tilde{\beta}x$, $\hat{x}^* = \frac{\tilde{\alpha}}{\tilde{\beta}}$ and $\sigma^{*2} = \frac{\tilde{\alpha}}{\tilde{\beta}}$ for the numerator. The approximate posterior mean is then

$$\hat{E}[\lambda|X^{(n)}] = \frac{\frac{\tilde{\alpha}^{1/2}}{\tilde{\beta}}\left(\frac{\tilde{\alpha}}{\tilde{\beta}}\right)^{\tilde{\alpha}}e^{-\tilde{\alpha}}}{\frac{(\tilde{\alpha}-1)^{1/2}}{\tilde{\beta}}\left(\frac{\tilde{\alpha}-1}{\tilde{\beta}}\right)^{\tilde{\alpha}-1}e^{-(\tilde{\alpha}-1)}}$$

6. Applications

In this section we present two applications of our results. First we approximate the posterior mean and variance of a Poisson mean parameter with a gamma prior distribution and compare the results to the exact values. In the second example we compute posterior means, variances and marginal densities of the three parameters of the Pareto model of Turnbull, Brown and Hu (1979) for the Stanford heart transplant data. These approximations are compared to numerical results obtained by Naylor and Smith (1982).

We begin with the Poisson example with a posterior density proportional to $\lambda^{\alpha-1} e^{-(\beta+n)\lambda}$ which is $\lambda^{\alpha-1} e^{-\beta\lambda}$ where $\hat{\alpha} = \alpha + s$ and $\hat{\beta} = \beta + n$. Thus the posterior mean and variance are given by $\frac{\hat{\alpha}}{\hat{\beta}}$ and $\frac{\hat{\alpha}}{\hat{\beta}^2}$, respectively. For approximating the posterior mean of λ, set $L(\alpha) = (\hat{\alpha} - 1)\log x - \hat{\beta}x$, $\hat{x} = \frac{\hat{\alpha} - 1}{\hat{\beta}}$ and $\hat{\sigma}^2 = \frac{\hat{\alpha} - 1}{\hat{\beta}^2}$ for the denominator, and $L^*(x) = \hat{\alpha}\log x - \hat{\beta}x$, $\hat{x}^* = \frac{\hat{\alpha}}{\hat{\beta}}$ and $\hat{\sigma}^2 = \frac{\hat{\alpha}}{\hat{\beta}^2}$ for the numerator. The approximate posterior mean is then

$$E(\lambda|x^{(n)}) = \frac{\frac{\hat{\alpha}}{\hat{\beta}^2}^{1/2}\left(\frac{\hat{\alpha}}{e}\right)^{\hat{\alpha}} e^{-\hat{\alpha}}}{\frac{(\hat{\alpha}-1)}{\hat{\beta}}^{1/2}\left(\frac{\hat{\alpha}-1}{e}\right)^{\hat{\alpha}-1} e^{-(\hat{\alpha}-1)}}$$

$$= \frac{\tilde{\alpha}}{\tilde{\beta}} \left(\frac{\tilde{\alpha}}{\tilde{\alpha} - 1} \right)^{\tilde{\alpha}-1/2} e^{-1}$$

$$= E[\lambda | X^{(n)}] \left(\frac{\tilde{\alpha}}{\tilde{\alpha} - 1} \right)^{\tilde{\alpha}-1/2} e^{-1}$$

if $\tilde{\alpha} > 1$; if $\tilde{\alpha} \leq 1$ then the approximation is not applicable since in this case the denominator integrand $\lambda^{\tilde{\alpha}-1} e^{\tilde{\beta}x}$ does not have an interior maximum. Note that the relative error is independent of $\tilde{\beta} = \beta + n$ and thus of the sample size; it only depends on $\tilde{\alpha} = \alpha + \Sigma X_i$.

Similar calculations lead to the variance approximation

$$\hat{Var}(\lambda | X^{(n)}) = \frac{\tilde{\alpha}}{\tilde{\beta}^2} \left[\frac{(\tilde{\alpha}+1)^2}{2} \left(\frac{\tilde{\alpha}+1}{\tilde{\alpha}-1} \right)^{\tilde{\alpha}-1/2} - \tilde{\alpha} \left(\frac{\tilde{\alpha}}{\tilde{\alpha} - 1} \right)^{2\tilde{\alpha}-1} \right] e^{-2}$$

$$= Var(\lambda | X^{(n)}) \left[\frac{(\tilde{\alpha}+1)^2}{2} \left(\frac{\tilde{\alpha}+1}{\tilde{\alpha}-1} \right)^{\tilde{\alpha}-1/2} - \tilde{\alpha} \left(\frac{\tilde{\alpha}}{\tilde{\alpha}-1} \right)^{2\tilde{\alpha}-1} \right] e^{-2} .$$

Again, the realtive error is independent of $\tilde{\beta} = \beta + n$. Table 6.1 lists values of the ratios of the approximate to the correct results for various values of $\tilde{\alpha}$. As can be seen, if $\tilde{\alpha} = \alpha + \Sigma X_i \geq 2$, which would be the case in most applications, then the error in the approximations does not exceed 4%!

As our second example we consider a three parameter model used by Turnbull, Brown and Hu (1974) to describe data from the Stanford heart transplant program and refered to by them as the Pareto model. This

$\hat{\alpha}$	2	3	4	6	8	10
$\hat{E}[\lambda \mid X^{(n)}]/E[\lambda \mid X^{(n)}]$	1.0405	1.0138	1.0069	1.0028	1.0015	1.0009
$\hat{Var}(\lambda \mid X^{(n)})/Var(\lambda \mid X^{(n)})$	.99914	.99994	.99999	1.0000	1.0000	1.0000

Table 6.1: Ratios of approximate to exact posterior means and standard deviations for Poisson data with a gamma prior.

model, described in Section 4.3 of their paper, views individual patients in the nontransplant group as having exponential lifetimes with mean ϕ, where ϕ is itself a random variable drawn independently for each patient from a gamma distribution with density proportional to $\phi^{p-1}e^{-\lambda\phi}$. Patients in the transplant group have a similar distribution but with $\tau\phi$ in place of ϕ for the residual lifetime after the transplant. The resulting likelihood function of the three parameters τ, λ, p is

$$\prod_{i=1}^{n} \frac{p\lambda^p}{(\lambda+x_i)^{p+1}} \prod_{i=n+1}^{N} \left(\frac{\lambda}{\lambda+x_i}\right)^p \prod_{j=1}^{m} \frac{\tau p\lambda^p}{(\lambda+y_j+\tau z_j)^{p+1}} \prod_{j=m+1}^{M} \left(\frac{\lambda}{\lambda+y_j+\tau z_j}\right)^p ,$$

where the x_i are the survival times in days of the $N = 30$ nontransplant patients, $n = 26$ of whom died, and y_j, z_j are the times to transplant and survival times after transplant, respectively, of the $M = 52$ transplant patients, $m = 34$ of whom died.

Naylor and Smith (1982) use this model with an improper uniform prior on the parameters τ, λ, p to illustrate their computational approach based on Gauss-Hermite quadrature, and we use the same improper prior distribution for the present illustration. Naylor and Smith point out the possibility of

Table 6.1. Ratios of approximate to exact posterior means and standard deviations for Poisson data with a gamma prior.

model, described in Section 4.3 of their paper, views individual lifetimes in the nontransplant group as having exponential lifetimes with mean θ_i, where ρ is itself a random variable drawn independently for each patient from a gamma distribution with density proportional to $\rho^{\gamma-1}e^{-\lambda\rho}$. Patients in the transplant group have a similar distribution but with $\theta_i\lambda$ in place of θ for the residual lifetime after the transplant. The resulting likelihood function of the three parameters τ, λ, ρ is

$$\prod_{i=1}^{n}\frac{\rho\lambda^{\rho}}{\Gamma(\rho)}(\lambda+x_i)^{\rho-1}\prod_{i=n+1}^{N}\frac{1}{\Gamma(\rho)}\left(\frac{\lambda}{\lambda+x_i}\right)^{\rho}\prod_{j=1}^{m}\frac{\rho\lambda^{\rho}}{\Gamma(\rho)(\lambda+y_j-\tau)^{\rho+1}}\prod_{j=m+1}^{M}\left(\frac{\lambda}{\lambda+y_j-\tau}\right)^{\rho}$$

where the x_i are the survival times in days of the $N = 30$ nontransplant patients, $n = 26$ of whom died, and y_j, z_j are the times to transplant and survival times after transplant, respectively, of the $M = 34$ transplant patients, $m = 24$ of whom died.

Tsiatis and Smith (1980) use this model with an improper uniform prior on the parameters τ, λ, ρ to illustrate their computational approach based on Gaussian-Hermite quadrature, and we use the same improper prior distribution for the present illustration. Naylor and Smith point out the possibility of

integrating out the parameter p analytically, but, following their example, we have chosen not to do this and to apply our approximations directly.

Table 6.2 lists the posterior means and standard deviations computed by the Laplace approximation method and by Naylor and Smith using Gauss-Hermite integration applied to an orthogonalized reparameterization. In addition, we list the posterior means and variances

	Posterior Means			Posterior Standard Deviation		
	τ	λ	p	τ	λ	p
Laplace	1.044	32.11	0.4926	.4944	16.09	.1381
Naylor & Smith	1.04	32.5	0.50	0.47	16.2	0.14
Marginal Laplace	1.048	32.59	0.4980	.4813	16.04	.1399
Marginal Gauss-Hermite (60-20-20)	1.042	32.40	0.4956	.4808	15.99	.1394
Monte Carlo (Importance sampling)	--	--	.5011	--	--	.1461

Table 6.2 Posterior Means and Standard Deviations for the Pareto Model.

obtained by integrating the Laplace and the Gauss-Hermite approximations to the marginal densities described below. These integrals were computed by a simple rectangular integration formula based on the 60 equally spaced points at which the marginals were evaluated. Finally, Table 6.2 shows the result of a Monte Carlo integration for the parameter p. These values were computed using importance sampling with 10,000 replications for each of the integrals $\int e^{l(\theta)}d\theta$, $\int pe^{l(\theta)}d\theta$ and $\int p^2e^{l(\theta)}d\theta$. We used normal importance weight functions with

independently and the parameter ρ analytically but, following their course, we have chosen not to do this and to apply our approximations directly.

Table 6.2 lists the posterior means and standard deviations computed by the Laplace approximation method and by Naylor and Smith using Gauss-Hermite integration applied to an augmented reparameterization. In addition, we list the posterior means and variances

	Posterior Means			Posterior Standard Deviation		
	γ	λ	ρ	γ	λ	ρ
Laplace	1.044	32.11	0.4756	.4946	11.09	.1321
Naylor & Smith	1.04	32.5	0.440	[illegible]	[illegible]	0.14
Numerical Laplace	1.043	32.39	0.4750	.4812	10.94	.1377
Numerical Gauss-Hermite (40-20-20)	1.042	32.40	0.4754	.4808	10.99	.1374
Monte Carlo (Importance sampling)	---	---	.5011		---	.1461

Table 6.2 Posterior Means and Standard Deviations for the Pareto Model.

obtained by integrating the Laplace and the Gauss-Hermite approximations to the marginal densities described below. These marginals were computed by a simple rectangular integration rule based on the 50 equally spaced points at which the marginals were evaluated. Finally, Table 6.2 shows the result of a Monte Carlo integration for the parameter ρ. These values were computed using importance sampling with 10,000 replications to each of the integrals $p_t(\theta)$, $p_{\gamma,\lambda}(\theta)$ and $p_{\gamma,\lambda,\rho}(\theta)$. The second normal importance weight functions with

means and covariances given by $(\hat{\theta}, \Sigma)$, $(\hat{\theta}^*, \Sigma^*)$, etc. Thus the Monte Carlo integration was used to correct the Laplace approximations to these integrals. As can be seen from this table, the largest relative difference between any of the Laplace approximations and the results of Naylor and Smith or the results obtained from the Gauss-Hermite marginals described below is around 4%.

In approximating the marginal densities, we selected a set of 60 equally spaced points for each parameter and then at each point computed approximations to the marginal densities by formula (5.1). A simple rectangular integration of these approximate densities produced integrals of approximately 1.2 in all three cases; thus renormalization was necessary. We then obtained plots of spline interpolations of the renormalized densities. We used 60 points for increased accuracy; however, 30 points produced identical pictures. In performing the maximizations for the individual grid points we proceeded outward from the MLE's, using each current set of optimal values as the starting values for the next maximization.

As a basis for comparison, for each of the 60 grid points selected for a given parameter we orthogonalized the other two parameters using the Σ matrix computed for the Laplace approximation at that point. We then integrated with respect to each of the two orthogonalized parameters using a 20 point Gauss-Hermite quadrature. The results were renormalized using a rectangular integration formula and plots were obtained. In all three cases the resulting plots were indistinguishable from the

means and covariances given by (θ_1, Σ_1), (θ_2, Σ_2^*), etc. Thus the Monte Carlo integration was used to correct the Laplace approximations to these integrals. As can be seen from this table, the largest relative difference between any of the Laplace approximations and the results of Naylor and Smith or the results obtained from the Gauss-Hermite marginals described below is around 4%.

To approximating the marginal densities, we selected a set of 60 equally spaced points for each parameter and then at each point computed approximations to the marginal densities by formula (5.1). A simple rectangular integration of these approximate densities produced integrals of approximately 1.2 in all three cases; thus renormalisation was necessary. We then obtained plots of spline interpolations of the renormalised densities. We used 60 points for increased accuracy; however, 30 points produced identical pictures. In performing the maximisations for the individual grid points we proceeded outward from the MLE's, using each current set of optimal values as the starting value for the next maximisation.

As a basis for comparison, for each of the 60 grid points evaluated for a given parameter we reoptimised the other two parameters using the Σ matrix computed for the Laplace approximation at that point. We then integrated with respect to each of the two orthogonalised parameters using a 20 point Gauss-Hermite quadrature. The results were renormalised using a similar integration formula and plots were obtained. In all three cases the resulting plots were indistinguishable from the

renormalized Laplace approximations. Figure 6.1 shows the plots for the marginal density of p together with a plot of the ratio of the renormalized Laplace approximation to the Gauss-Hermite calculation. In Figure 6.2 we show the results for all three parameters. The solid lines are the superpositions of the renormalized Laplace approximation and the Gauss-Hermite calculations; the broken lines are the asymptotic normal approximations, which have been included as a basis for comparison. As an aside, Naylor and Smith show a more peaked asymptotic normal distribution for λ based on the estimated asymptotic standard deviation of 6 reported in Turnbull, Brown and Hu. According to our calculations the estimated asymptotic standard deviation of λ is in fact equal to 10.25.

In addition to considering the accuracy of the Laplace approximations, it is also worth noting the relative computing time requirements. The computations were performed on the University of Minnesota's Cyber 730 computer, which is a relatively fast machine. A rough guess is that the CPU requirements on a VAX 750 with floating point hardware would be roughly 5 times higher. Table 6.3 lists the CPU time required by the Laplace moment calculations for 30 and 60 grid points, the Gauss-Hermite calculations for 30 grid and 10 Gauss-Hermite points and for 60 grid and 20 Gauss-Hermite points. Finally we give the time required by the Monte Carlo integration.

	CPU Time	Real Time
Laplace Moments	.14 sec	2 - 5 sec
Laplace marginal 30 points	1 sec	7 - 10 sec
Laplace marginal 60 points	1.9 sec	8 - 15 sec
Gauss-Hermite 30-10-10	10 sec	50 - 60 sec
Gauss-Hermite 60-20-20	69.8 sec	4 - 5 min
Monte Carlo 10,000 points	164 sec	18 min

Table 6.3 Computing Times for the Pareto Model

All computations were done in the interactive FORTRAN system. Therefore, in addition to the CPU requirements Table 6.3 also shows the real time requirements, the time between issuing the final instructions and the beginning of the printing of the results. To evaluate the usefulness of a technique as a tool for interactive statistical analysis these times, though they are highly system dependent, may be more relevant than the CPU times.

The most striking feature is that the Monte Carlo computation of the moments of p required nearly three minutes of CPU time and 18 minutes of real time. Thus we chose not to use this approach for any further moments or for approximating the marginal distributions. The Gauss-Hermite calculations using 30 grid points and 10 Gauss-Hermite points took a little over a minute of real time and eight seconds of CPU time. But it should be remembered that these figures would be multiplied by a factor of

10 for each additional parameter. The computations for the Laplace approximations on the other hand were virtually instantaneous, making it possible to look at moments and marginal plots for several choices of prior distributions in quick succession.

Gauss-Hermite marginal calculations using a single orthogonalization based on the asymptotic or the posterior covariance matrix appear to be less accurate than ones based on the adaptive orthogonalizations used here. In a calculation of the marginal density of p using a single orthogonalization there was a noticeable difference between the results based on 10 and 20 Gauss-Hermite points. The results using 20 points with a single orthogonalization were, however, very close to results of adaptive orthogonalizations using 10 or 20 points. The two curves using adaptive orghogonalizations with 10 and 20 Gauss-Hermite points, on the other hand, were indistinguishable from one another. Thus the adaptive approach, which can be thought of as computing corrections starting from the Laplace approximations, appears to be more accurate.

7. Concluding Remarks

In this paper we have introduced and studied a new approximation technique for posterior moments, predictive distributions and marginal posterior distributions. While the formal error analyses we have presented are all asymptotic, in all cases we have tried so far the approximations perform extremely well even for relatively small sample sizes. The approximations are very easy to use, requiring essentially only the ability to compute maximum likelihood estimates and second derivatives of log-likelihood functions, and the computing time requirements are generally minimal. Thus they may go a long way towards facilitating the use of Bayesian methods in interactive data analyses.

Several open questions remain to be investigated. One is to determine the exact conditions under which the approximate marginalization approach produces exact results. As mentioned above, it may be possible to adapt the method used by Daniels (1980) for the saddle point approximation to the present setting. Another interesting problem would be to determine whether the approximations proposed here remain accurate when numerical derivatives are used in place of analytic ones in cases where closed form derivatives are not available. It would also be desirable to find a more satisfactory approach than the translation method mentioned at the end of Section 3 for approximating the posterior moments of parameters taking on both positive and negative values.

The approximations of this paper may also prove helpful in

certain theoretical problems such as developing tractable Bayesian approaches to log-linear models and to experimental design for nonlinear models. An extension of our methods that is currently being explored is to multimodal posteriors such as the poly-t distributions of Dickey (1968) and Drèze (1977).

In conclusion, we would like to emphasise that we do not think of these approximations as replacements for exact calculations in situations where extremely accurate results are needed. Instead, we consider them to be simple first approximations that are easy to obtain, are often sufficiently accurate in their own right, and generally provide good starting points for exact computations, should these be required.

Acknowledgements

We would like to thank Rob Kass for many helpful conversations and for bringing the results of Leonard and of Mostellar and Wallace to our attention.

Appendix 1

This appendix presents the details of the derivation of the constant a in equation (2.3). Set $u_i = n^{1/2}(\theta_i - \hat{\theta}_i)$, $u = (u_1, \ldots, u_p)$, and let $U_1, \ldots, U_p$ be jointly normal random variables with mean zero and covariance matrix Σ. Then

$$n(L(\theta)-L(\hat{\theta}) + \frac{1}{2}u^T\Sigma^{-1}u) = \frac{1}{6n^{1/2}} \sum_{ijk} u_i u_j u_k L_{ijk} + \frac{1}{24n} \sum_{ijkl} u_i u_j u_k u_l L_{ijkl} + \frac{1}{120n^{3/2}} \sum_{ijklm} u_i u_j u_k u_l u_m L_{ijklm} + R,$$

where R is a higher-order error term. Thus

$$\begin{aligned} H(u) &= \exp\{n(L(\theta)-L(\hat{\theta}) + \frac{1}{2} u^T\Sigma u^{-1})\} \\ &= 1 + \frac{1}{6n^{1/2}} \sum_{ijk} u_i u_j u_k L_{ijk} + \frac{1}{24n} \sum_{ijkl} u_i u_j u_k u_l L_{ijkl} \\ &\quad + \frac{1}{120n^{3/2}} \sum_{ijklm} u_i u_j u_k u_l u_m L_{ijklm} \\ &\quad + \frac{1}{2}\left(\frac{1}{6n^{1/2}} \sum_{ijk} u_i u_j u_k L_{ijk} + \frac{1}{24n} \sum_{ijkl} u_i u_j u_k u_l L_{ijkl}\right)^2 + R' \\ &= 1 + \frac{1}{6n^{1/2}} \sum_{ijk} u_i u_j u_k L_{ijk} + \frac{1}{24n} \sum_{ijkl} u_i u_j u_k u_l L_{ijkl} \\ &\quad + \frac{1}{120n^{3/2}} \sum_{ijklm} u_i u_j u_k u_l u_m L_{ijklm} \\ &\quad + \frac{1}{72n} \sum_{ijklmr} u_i u_j u_k u_l u_m u_r L_{ijk} L_{lmr} \\ &\quad + \frac{1}{72n^{3/2}} \sum_{ijklmrs} u_i u_j u_k u_l u_m u_r u_s L_{ijk} L_{lmrs} + R''. \end{aligned}$$

Now the integral to be approximated can be written as

$$\int e^{nL(\theta)}d\theta = \exp\{nL(\hat{\theta})\} \int H(u)e^{(1/2)u^T \Sigma^{-1} u}du$$

$$= (2\pi/n)^{P/2}(\det \Sigma)^{1/2} e^{nL(\hat{\theta})}$$

$$\cdot \left[(2\pi)^{-P/2}(\det \Sigma)^{-1/2} \int H(u)e^{(-1/2)u^T \Sigma^{-1} u}du\right]$$

$$= (2\pi/n)^{P/2}(\det \Sigma)^{1/2} e^{nL(\hat{\theta})} E[H(U)].$$

Finally, using the fact that odd order moments of multivariate normal vectors vanish, we have

$$E[H(U)] = 1 + \frac{1}{n}\left(\frac{1}{24} \sum_{ijkl} E[U_i U_j U_k U_l] L_{ijkl} + \frac{1}{72} \sum_{ijklmr} E[U_i U_j U_k U_l U_m U_r] L_{ijk} L_{lmr}\right) + O(n^{-2}).$$

The details of the expression for a in (2.3) now follow by computing the expectations $E[U_i U_j U_k U_l]$ and $E[U_i U_j U_k U_l U_m U_r]$; this can be done by differentiating the moment generating function.

Appendix 2

To illustrate how the statements of this paper can be made precise, this appendix gives more refined statements and proofs of approximations (2.1), (2.4) and (3.1). Similar refinements can be given for the multiparameter results (2.3) and (2.5) and the marginal approximation (5.2); these are omitted. We begin with (2.1).

Theorem 1.

Let Θ be an open set, and let L, ℓ_n and v be real valued functions on Θ. Let $\Theta_+ = \{\theta: v(\theta) > 0\} \neq \emptyset$, let $V = \log v$ and $L_n = \ell_n + V/n$ on Θ_+, and assume that L, ℓ_n and v satisfy the following five conditions. (In this appendix the log-likelihood will be represented by $n\ell_n$)

(i) ℓ, ℓ_n and v are 8 times continuously differentiable.

(ii) L has a unique maximum at $\hat{\theta}$, $L''(\hat{\theta}) < 0$, and $v(\hat{\theta}) > 0$.

(iii) $\ell_n(\hat{\theta}) \to L(\hat{\theta})$, $(\frac{d}{d\theta})^k \ell_n(\hat{\theta}) \to (\frac{d}{d\theta})^k L(\hat{\theta})$ for $k \leq 7$ and $(\frac{d}{d\theta})^8 \ell_n \to (\frac{d}{d\theta})^k L$ uniformly on some neighborhood of $\hat{\theta}$.

(iv) v is bounded above and $\int v(\theta)d\theta < \infty$.

(v) For all small δ there is a $c(\delta) > 0$ such that $\ell_n(\theta) - L(\hat{\theta}) \leq -c(\delta) < 0$ for all θ with $|\theta-\hat{\theta}| \geq \delta$ once n is large enough.

Then for all large n, $ve^{n\ell_n}$ has a unique maximum at a point $\hat{\theta}_n \in \Theta_+$ and $\hat{\theta}_n \to \hat{\theta}$. Furthermore, if $\sigma_n^2 = -1/L''(\hat{\theta}_n)$,

$$\mu_k = \frac{k!}{(k/2)!\,2^{k/2}} \text{ for even } k, \quad h_{n,k} = \frac{\sigma_n^k}{k!}\left(\frac{d}{d\theta}\right)^k L_n(\hat{\theta}_n),$$

$$a_n = \mu_4 h_{n,4} + \frac{1}{2}\mu_6 h_{n,3}^2$$

and

$$b_n = \mu_6 h_{n,6} + \frac{1}{2}\mu_8 h_{n,4}^2 + \mu_8 h_{n,3} h_{n,5} + \frac{1}{2}\mu_{10} h_{n,3}^2 h_{n,4} + \frac{1}{24}\mu_{12} h_{n,3}^4,$$

then a_n and b_n are bounded and

$$\int_{\Theta} v(\theta) e^{n\ell_n(\theta)}\,d\theta = \sqrt{2\pi}\,\sigma_n n^{-1/2} \exp\{nL_n(\hat{\theta}_n)\}\left(1 + \frac{a_n}{n} + \frac{b_n}{n^2} + O(n^{-3})\right)$$

Proof

The proof of the first claim is essentially identical to the proof of the consistency of the MLE as given, for example, on pages 83-84 of Walker (1969):

Since $\ell_n(\hat{\theta}) \to L(\hat{\theta})$ and v is bounded, property (v) implies that for any small $\delta > 0$ once n is large enough L_n has a maximum and that maximum is only attained in the set $\{\theta: |\theta-\hat{\theta}| < \delta\}$. (It is assumed that δ is small enough such that $\{\theta: |\theta-\hat{\theta}| < \delta\} \subseteq \Theta_+$). Since L_n is differentiable, any maximum on this set must solve $L_n'(\theta) = 0$. Since $L''(\hat{\theta}) < 0$, L'' is continuous, and $L_n'' \to L''$ unformly on some neighborhood of $\hat{\theta}$ (this follows from property (iii)), the equation $L_n'(\theta) = 0$ cannot have more than one solution. Thus for large n, L_n has a unique maximum at a point $\hat{\theta}_n$. Since for any small δ, $|\hat{\theta}_n - \hat{\theta}| < \delta$ for all large n, we have

$\hat{\theta}_n \to \hat{\theta}$. This proves the first claim.

Once n is large enough for $\hat{\theta}_n$ to exist and σ_n^2 to be positive, we can write

$$\int_{\Theta} v(\theta)\, e^{n\mathcal{I}_n(\theta)}\, d\theta = \sqrt{2\pi}\, \sigma_n n^{-1/2} \exp\{nL_n(\hat{\theta}_n)\} \cdot \left[\sqrt{\frac{n}{2\pi\sigma_n^2}} \int_{\Theta} v(\theta)\, e^{(n\mathcal{I}_n(\theta) - L_n(\hat{\theta}_n))}\, d\theta\right]$$

Thus to prove the second claim we need to show that the term in brackets is equal to $(1 + \frac{a_n}{n} + \frac{b_n}{n^2} + O(n^{-3}))$. Boundedness of a_n and b_n follows from (iii) and the fact that $\hat{\theta}_n \to \hat{\theta}$. Choose $\delta > 0$ small enough for (v) to hold and for $\{\theta: |\theta - \hat{\theta}| \le \delta\} \subseteq \Theta_+$. Since $\hat{\theta}_n \to \hat{\theta}$ and $L_n \to L$ uniformly near $\hat{\theta}$, it follows that for all large n, $\mathcal{I}_n(\theta) - L_n(\hat{\theta}_n) \le -c(\delta)/2$ when $|\theta - \hat{\theta}_n| \le \delta/2$. Hence

$$\begin{aligned}
&\sqrt{\frac{n}{2\pi\sigma_n^2}} \left| \int_{\Theta - \{\theta: |\theta - \hat{\theta}_n| > \delta/2\}} v(\theta) e^{n(\mathcal{I}_n(\theta) - L_n(\hat{\theta}_n))}\, d\theta \right| \\
&\quad \le \sqrt{\frac{n}{2\pi\sigma_n^2}} \int_{\Theta} v(\theta) e^{-nc(\delta)/2}\, d\theta \\
&\quad \le \sqrt{\frac{n}{2\pi\sigma_n^2}}\, e^{-nc(\delta)/2} \int_{\Theta} v(\theta)\, d\theta \\
&\quad = O(n^{-3}) ,
\end{aligned}$$

and thus

$$\sqrt{\frac{n}{2\pi\sigma_n^2}} \int_{\Theta} v(\theta) e^{n(\mathcal{I}_n(\theta) - L_n(\hat{\theta}_n))}\, d\theta$$

$$= \sqrt{\frac{n}{2\pi\sigma_n^2}} \int_{|\theta-\hat{\theta}_n|\le\delta/2} v(\theta) e^{n(L_n(\theta)-L_n(\hat{\theta}_n))} d\theta + O(n^{-3})$$

$$= \sqrt{\frac{n}{2\pi\sigma_n^2}} \int_{|\theta-\hat{\theta}_n|\le\delta/2} e^{n(L_n(\theta)-L_n(\hat{\theta}_n))} d\theta + O(n^{-3}) .$$

Now

$$L_n(\theta) - L_n(\hat{\theta}_n) = \frac{1}{2} (\theta - \hat{\theta}_n) L_n''(\xi_n)$$

for some ξ_n between $\hat{\theta}_n$ and θ. Since $L_n'' \to L''$ uniformly near $\hat{\theta}$, L'' is continuous and $L''(\hat{\theta}) < 0$, for small enough δ we have

$$L_n(\theta) - L_n(\hat{\theta}_n) \le \frac{1}{4} (\theta - \hat{\theta}_n)^2 L''(\hat{\theta})$$

for all θ with $|\theta - \hat{\theta}_n| \le \delta/2$ once n is large enough. Thus for any small $\delta > 0$ and any α with $0 < \alpha < 1/2$

$$\sqrt{\frac{n}{2\pi\sigma_n^2}} \int_{n^{-\alpha}\le|\theta-\hat{\theta}_n|\le\delta/2} v(\theta) e^{n(L_n(\theta)-L_n(\hat{\theta}_n))} d\theta$$

$$\le \sqrt{\frac{n}{2\pi\sigma_n^2}} \int_{n^{-\alpha}\le|\theta-\hat{\theta}_n|\le\delta/2} e^{\frac{1}{4} n(\theta-\hat{\theta}_n)^2 L''(\hat{\theta})} d\theta$$

$$= O(n^{-3}) ,$$

and therefore,

$$\sqrt{\frac{n}{2\pi\sigma_n^2}} \int_H v(\theta) e^{n(L_n(\theta)-L_n(\hat{\theta}_n))} d\theta$$

$$= \sqrt{\frac{n}{2\pi\sigma_n^2}} \int_{|\theta-\hat{\theta}_n|\le n^{-\alpha}} e^{n(L_n(\theta)-L_n(\hat{\theta}))} d\theta + O(n^{-3})$$

for any α with $0 < \alpha < 1/2$.

By Taylor's theorem, setting $z = \sqrt{n}(\theta-\hat{\theta})/\sigma_n$ we can write

$$H_n(\theta) = n(L_n(\theta)-L_n(\hat{\theta}_n) + \frac{(\theta-\hat{\theta}_n)^2}{2\sigma n^2})$$

$$= \frac{z^3}{n^{1/2}} h_{n,3} + \ldots + \frac{z^7}{n^{5/2}} h_{n,7} + \frac{z^8}{n^3} g_n(z)$$

$$= \tilde{H}_n(z) + \frac{z^8}{n^3} g_n(z) ,$$

where $g_n(z) = \frac{\sigma_n^8}{8!} L_n^{(viii)}(\xi_n)$ for some ξ_n between θ and $\hat{\theta}_n$. Since $L_n^{(viii)} \to L^{(viii)}$ uniformly near $\hat{\theta}$ and $L^{(viii)}$ is continuous, the sequence $\sup\{|g_n(z)|: |z| \le n^{1/2-\alpha}/\sigma_n\}$ is bounded. This implies that $\sup\{|H_n(\theta)|: |\theta-\hat{\theta}_n| \le n^{-\alpha}\} \to 0$ if $\alpha > 1/8$. Hence we have

$$\exp\{H_n(\theta)\} = \sum_{k=0}^{5} \frac{1}{k!} \tilde{H}_n(z)^k + \frac{1}{n^3} q_n(z)$$

$$= G_n(z) + \frac{1}{n^3} q_n(z)$$

for $|z| \le n^{1/2-\alpha}/\sigma_n$ and $1/8 < \alpha < 1/2$, where the functions $q_n(z)$ are bounded in absolute value by a polynomial $P(z)$ that does not depend on n. Thus for $1/8 < \alpha < 1/2$

$$\sqrt{\frac{n}{2\pi\sigma n^2}} \int_H v(\theta) e^{n(L_n(\theta)-L_n(\hat{\theta}_n))} d\theta$$

$$= \frac{1}{\sqrt{2\pi}} \int_{|z| \le n^{1/2-\alpha}/\sigma_n} (G_n(z) + \frac{1}{n^3} q_n(z)) e^{-z^2/2} dz + O(n^{-3})$$

$$= \frac{1}{\sqrt{2\pi}} \int_{|z| \le n^{1/2-\alpha}/\sigma_n} G_n(z) e^{-z^2/2} dz + O(n^{-3})$$

$$= \frac{1}{\sqrt{2\pi}} \int G_n(z) d^{-z^2/2} dz + O(n^{-3}) \ .$$

Finally, using the fact that

$$\frac{1}{\sqrt{2\pi}} \int z^k e^{-z^2/2} dz = \begin{cases} \mu_k & \text{if } k \text{ is even} \\ 0 & \text{if } k \text{ is odd} \end{cases}$$

we have

$$\frac{1}{\sqrt{2\pi}} \int e^{-z^2/2} G_n(z) dz$$

$$= \frac{1}{\sqrt{2\pi}} \int e^{-z^2/2} [1 + \frac{z^3}{n^{1/2}} h_{n,3} + \frac{z^4}{n} h_{n,4} + \frac{z^5}{n^{3/2}} h_{n,5} + \frac{z^6}{n^2} h_{n,6} + \frac{z^7}{n^{5/2}} h_{n,7}$$

$$+ \frac{1}{2} (\frac{z^6}{n} h_{n,3}^2 + \frac{z^8}{n^2} h_{n,4}^2 + \frac{2z^7}{n^{3/2}} h_{n,3} h_{n,4} + \frac{2z^8}{n^2} h_{n,3} h_{n,5}$$

$$+ \frac{2z^9}{n^{5/2}} h_{n,3} h_{n,6} + \frac{2z^9}{n^{5/2}} h_{n,4} h_{n,5})$$

$$+ \frac{1}{6} (\frac{z^9}{n^{3/2}} h_{n,3}^3 + \frac{3z^{10}}{n^2} h_{n,3}^2 h_{n,4} + \frac{3z^{11}}{n^{5/2}} h_{n,3}^2 h_{n,4} + \frac{3z^{11}}{n^{5/2}} h_{n,3} h_{n,4}^2)$$

$$+ \frac{1}{24} (\frac{z^{12}}{n^2} h_{n,3}^4 + \frac{4z^{13}}{n^{5/2}} h_{n,3}^3 h_{n,4})] dz + O(n^{-3})$$

$$= 1 + \frac{1}{n} \mu_4 h_{n,4} + \frac{1}{n^2} \mu_6 h_{n,6} + \frac{1}{2n} \mu_6 h_{n,3}^2 + \frac{1}{2n^2} \mu_8 h_{n,4}^2$$

$$+ \frac{1}{n^2} \mu_8 h_{n,3} h_{n,5} + \frac{1}{2n^2} \mu_{10} h_{n,3}^2 h_{n,4} + \frac{1}{24n^2} \mu_{12} h_{n,3}^4 + O(n^{-3})$$

$$= 1 + \frac{1}{n}\left(\mu_4 h_{n,4} + \frac{1}{2}\mu_6 h_{n,3}^2\right)$$
$$+ \frac{1}{n^2}\left(\mu_6 h_{n,6} + \frac{1}{2}\mu_8 h_{n,4}^2 + \mu_8 L_{n,3} L_{n,5} + \frac{1}{2}\mu_{10} h_{n,3}^2 h_{n,4}\right.$$
$$\left. + \frac{1}{24}\mu_{12} h_{n,3}^4\right) + O(n^{-3}). \qquad \square$$

Next we consider (2.4).

Theorem 2.

Let Θ, L, ℓ_n and v satisfy the assumptions of Theorem 1. Let w be a real valued 8 times continuously differentiable function on Θ, set $\Theta_+^* = \{\theta: w(\theta), v(\theta) > 0\}$, and set $W = \log w$ and $L_n^* = \ell_n + (V + W)/n = L_n + W/n$ on Θ_+^*. Assume that $w(\hat{\theta}) > 0$, vw is bounded above, and $\int_\Theta w(\theta)v(\theta)d\theta < \infty$.

Using the notation $f^{(k)}(x) = \left(\frac{d}{dx}\right)^k f(x)$, set $W_{n,k} = W^{(k)}(\hat{\theta}_n)$. Then for all large n, $wve^{n\ell_n}$ has a unique maximum at a point $\hat{\theta}_n^* \in \Theta_+^*$, $\hat{\theta}_n^* \to \hat{\theta}$, and

$$\text{(A.1)} \qquad \hat{\theta}_n^* = \hat{\theta}_n + W_{n,1}\sigma_n^2 \frac{1}{n} + O(n^{-3}).$$

If $\sigma_n^{*2} = -1/L_n^{*(2)}(\hat{\theta}_n^*)$, $h_{n,k}^* = \frac{1}{k!}\sigma_n^{*k} L_n^{*(k)}(\hat{\theta}_n^*)$,

$$a_n^* = \mu_4 h_{n,4}^* + \frac{1}{2}\mu_6 h_{n,3}^{*2}$$

and

$$b_n^* = \mu_6 h_{n,3}^{*2} + \frac{1}{2}\mu_8 h_{n,4}^{*2} + \mu_8 h_{n,3}^* h_{n,5}^* + \frac{1}{2}\mu_{10} h_{n,3}^{*2} h_{n,4}^* + \frac{1}{24}\mu_{12} h_{n,3}^{*4},$$

then

(A.2) $$\int_{\Theta} w(\theta)v(\theta)e^{n\mathfrak{L}_n(\theta)}d\theta$$

$$= \sqrt{2\pi}\,\sigma_n^{*}n^{-1/2}\exp\{nL_n^{*}(\hat{\theta}_n^{*})\}(1 + \frac{a_n^{*}}{n} + \frac{b_n^{*}}{n^2} + O(n^{-3})).$$

Finally, if $L_{n,k} = L_n^{(k)}(\hat{\theta}_n)$,

$$d_{n,1} = \frac{1}{24}\mu_4\sigma_n^{6}L_{n,5} + \frac{1}{12}\mu_4\sigma_n^{8}L_{n,3}L_{n,4} + \frac{1}{6}\mu_6h_{n,3}\sigma_n^{5}L_{n,4} + \frac{1}{4}\mu_6h_{n,3}\sigma_n^{7}L_{n,3}^{2}$$

$$d_{n,2} = \frac{1}{12}\mu_4\sigma_n^{6}L_{n,4} + \frac{1}{4}\mu_6h_{n,3}\sigma_n^{5}L_{n,3}$$

$$d_{n,3} = \frac{1}{6}\mu_6h_{n,3}\sigma_n^{3}$$

$$d_{n,4} = \frac{1}{24}\mu_4\sigma_n^{4},$$

and $c_n = W_{n,1}d_{n,1} + W_{n,2}d_{n,2} + W_{n,3}d_{n,3} + W_{n,4}d_{n,4}$, then c_n and $d_{n,1}, \ldots, d_{n,4}$ are bounded and

(A.3) $$\frac{\int_{\Theta} w(\theta)v(\theta)e^{n\mathfrak{L}_n(\theta)}d\theta}{\int_{\Theta} v(\theta)e^{n\mathfrak{L}_n(\theta)}d\theta} = \frac{\sigma_n^{*}}{\sigma_n}\exp\{n(L_n^{*}(\hat{\theta}_n^{*})-L_n(\hat{\theta}_n))\} \cdot (1 + \frac{c_n}{n^2} + O(n^{-3})).$$

Proof

First note that the functions L, $\mathfrak{L}_n$ and $v^{*} = wv$ satisfy conditions (i) - (v) of Theorem 1. Hence $\hat{\theta}_n^{*}$ exists and is unique for all large n, $\hat{\theta}_n^{*} \to \hat{\theta}$, and (A.2) holds. To prove (A.1), note that $\hat{\theta}_n^{*}$ and $\hat{\theta}_n$ must solve the equations $L_n^{*(1)}(\theta) = 0$ and $L_n^{(1)}(\theta) = 0$, respectively. Thus for large n we can write

$$0 = L_n^{*(1)}(\hat{\theta}_n^*) = L_n^{*(1)}(\hat{\theta}) + (\hat{\theta}_n^* - \hat{\theta}_n)L_n^{*(2)}(\xi_n)$$
$$= \frac{1}{n} W_{n,1} + L_n^{(1)}(\hat{\theta}_n) + (\hat{\theta}_n^* - \hat{\theta}_n)L_n^{*(2)}(\xi_n)$$
$$= \frac{1}{n} W_{n,1} + (\hat{\theta}_n^* - \hat{\theta}_n)L_n^{*(2)}(\xi_n)$$

for some ξ_n between $\hat{\theta}_n$ and $\hat{\theta}_n^*$. Since $\xi_n \to \hat{\theta}$, $L_n^{(2)}$ converges uniformly near $\hat{\theta}$ to the continuous function $L^{(2)}$, and $L^{(2)}(\hat{\theta}_n) < 0$, this implies that for large n

$$\hat{\theta}_n^* = \hat{\theta}_n - \frac{1}{n} W_{n,1}/L_n^{*(2)}(\xi_n) = O(n^{-1})$$

Thus $\xi_n - \hat{\theta}_n = O(n^{-1})$ as well, and by the uniform convergence of $L_n^{*(3)}$ near $\hat{\theta}$, this implies that

$$L_n^{*(2)}(\xi_n) = L_n^{(2)}(\hat{\theta}_n) + O(n^{-1}) = -1/\sigma_n^2 + O(n^{-1}) .$$

So

$$\hat{\theta}_n^* = \hat{\theta}_n + \frac{1}{n} W_{n,1}\sigma_n^2 + O(n^{-2}) ,$$

as claimed.

To prove (A.3), note that

$$\frac{\int_\Theta w(\theta)v(\theta)e^{n\hat{L}_n(\theta)}d\theta}{\int_\Theta v(\theta)e^{n\hat{L}_n(\theta)}d\theta} = \frac{\sigma_n^*}{\sigma_n} \exp\{n(L_n^*(\hat{\theta}_n^*)-L_n(\hat{\theta}_n))\}$$
$$\cdot \frac{(1 + \frac{a_n^*}{n} + \frac{b_n^*}{n^2} + O(n^{-3}))}{(1 + \frac{a_n}{n} + \frac{b_n}{n} + O(n^{-3}))}$$

$$= \frac{\sigma_n^*}{\sigma_n} \exp\{n(L_n^*(\hat{\theta}_n^*) - L_n(\hat{\theta}_n))\}$$

$$\cdot(1 + \frac{a_n^* - a_n}{n} + \frac{b_n^* - b_n - a_n(a_n^* - a_n)}{n^2} + O(n^{-3})) .$$

We will show that $a_n^* - a_n = \frac{c_n}{n} + O(n^{-2})$; a similar argument can be used to show that $b_n^* - b_n = O(n^{-1})$. Now, for $k \leq 7$

$$L_n^{*(k)}(\hat{\theta}_n^*) = L_n^{(k)}(\hat{\theta}_n^*) + \frac{1}{n} W^{(k)}(\hat{\theta}_n^*)$$

$$= L_{n,k} + (\hat{\theta}_n^* - \hat{\theta}_n)L_{n,k+1} + \frac{1}{n} W_{n,k} + O(n^{-2})$$

$$= L_{n,k} + \frac{1}{n}(W_{n,1}\sigma_n^2 L_{n,k+1} + W_{n,k}) + O(n^{-2}) .$$

As a result,

$$\sigma_n^{*2} = -1/L_n^{*(2)}(\hat{\theta}_n^*)$$

$$= -[L_{n,2} + (W_{n,1}\sigma_n^2 L_{n,3} + W_{n,2})\frac{1}{n} + O(n^{-2})]^{-1}$$

$$= \sigma_n^2 + \frac{1}{n}(W_{n,1}\sigma_n^2 L_{n,3} + W_{n,2})\sigma_n^4 + O(n^{-2}) .$$

Combining these two expansions, we have

$$h_{n,k}^* = \frac{1}{k!}\sigma_n^{*k} L_n^{*(k)}(\hat{\theta}_n^*)$$

$$= \frac{1}{k!}[\sigma_n^2 + \frac{1}{n}(W_{n,1}\sigma_n^2 L_{n,3} + W_{n,2})\sigma_n^4 + O(n^{-2})]^{k/2}$$

$$\cdot [L_{n,k} + \frac{1}{k}(W_{n,1}\sigma_n^2 L_{n,k+1} + W_{n,k}) + O(n^{-2})]$$

$$= h_{n,k} + \frac{1}{nk!}[\sigma_n^k(W_{n,1}\sigma_n^2 L_{n,k+1} + W_{n,k})$$

$$+ \frac{k}{2}L_{n,k}\sigma_n^{k+2}(W_{n,1}\sigma_n^2 L_{n,3} + W_{n,2})] + O(n^{-2})$$

$$= h_{n,k} + \frac{1}{nk!}[W_{n,1}(\sigma_n^{k+2}L_{n,k+1} + \frac{k}{2}\sigma_n^{4+k}L_{n,3}L_{n,k})$$

$$+ W_{n,2}\,\tfrac{k}{2}L_{n,k}\sigma_n^{2+k} + W_{n,k}\sigma_n^{k}] + O(n^{-2}) \ .$$

Thus

$$h_{n,4}^{*} - h_{n,4} = \tfrac{1}{24n}[W_{n,1}(\sigma_n^{6}L_{n,5} + 2\sigma_n^{8}L_{n,3}L_{n,4}) + W_{n,2}2\sigma_n^{6}L_{n,4} + W_{n,4}\sigma_n^{4}] + O(n^{-2})$$

and

$$h_{n,3}^{*2} - h_{n,3}^{2} = 2h_{n,3}(h_{n,3}^{*} - h_{n,3}) + O(n^{-2})$$
$$= \tfrac{1}{3n}h_{n,3}[W_{n,1}(\sigma_n^{5}L_{n,4} + \tfrac{3}{2}\sigma_n^{7}L_{n,3}^{2}) + W_{n,2}\,\tfrac{3}{2}\sigma_n^{5}L_{n,3} + W_{n,3}\sigma_n^{3}] + O(n^{-2}) \ .$$

Finally

$$a_n^{*} - a_n = \mu_4(h_{n,4}^{*} - h_{n,4}) + \tfrac{1}{2}\mu_6(h_{n,3}^{*2} - h_{n,3}^{2})$$
$$= \tfrac{1}{n}[W_{n,1}(\tfrac{1}{24}\mu_4\sigma_n^{6}L_{n,5} + \tfrac{1}{12}\mu_4\sigma_n^{8}L_{n,3} + \tfrac{1}{6}\mu_6h_{n,3}\sigma_n^{5}L_{n,4} + \tfrac{1}{4}\mu_6h_{n,3}\sigma_n^{7}L_{n,3}^{2}) + W_{n,2}(\tfrac{1}{12}\mu_4\sigma_n^{6}L_{n,4} + \tfrac{1}{4}\mu_6h_{n,3}\sigma_n^{5}L_{n,3}) + W_{n,3}\,\tfrac{1}{6}\mu_6h_{n,3}\sigma_n^{3} + W_{n,4}\,\tfrac{1}{24}\mu_4\sigma_n^{4}] + O(n^{-2})$$
$$= \tfrac{1}{n}[W_{n,1}d_{n,1} + W_{n,2}d_{n,2} + W_{n,3}d_{n,3} + W_{n,4}d_{n,4}] + O(n^{-2})$$
$$= \frac{c_n}{n} + O(n^{-2}) \ .$$

Boundedness of the coefficients $d_{n,1}, \ldots, d_{n,4}$ and c_n follows from the convergence of $\hat{\theta}_n$ to $\hat{\theta}$ and the uniform convergence near $\hat{\theta}$ of $L_n^{(k)}$ for $k \le 7$. □

If L_n and L_n^* are defined as in (3.1), then, under mild regularity conditions, the assumptions of the previous two theorems are satisfied for almost all sample sequences, and thus (3.1) holds almost surely. One possible set of regularity conditions, adapted from the conditions used by Walker (1969), is given in the following theorem.

Theorem 3.

Let $\{f(x|\theta): \theta \in \Theta\}$ be a family of densities with respect to a σ-finite measure μ, fix a $\theta_0 \in \Theta$, and let $f_0(x) = f(x|\theta_0)$. Let X_1, X_2, ... be a sequence of i.i.d. random variables with density f_0. Assume that the following regularity conditions are satisfied.

(A1) Θ is an open subset of the real line.

(A2) The set of points $\{x: f(x|\theta) > 0\}$ is independent of θ and is denoted by $\mathcal{X}$.

(A3) If θ_1, θ_2 are two distinct points of Θ, then

$$\mu\{x: f(x|\theta_1) \neq f(x|\theta_2)\} > 0,$$

i.e. the distributions of X_i given $\theta = \theta_1$ and $\theta = \theta_2$ are different.

(A4) Let $x \in \mathcal{X}$, $\theta' \in \Theta$. Then for all θ such that $|\theta-\theta'| < \delta$, with δ sufficiently small,

$$|\log f(x|\theta) - \log (x|\theta')| < H_\delta(x,\theta') ,$$

where $\lim_{\delta \to 0} H_\delta(x,\theta') = 0$, and

$$\lim_{\delta \to 0} \int_{\mathcal{X}} H_\delta(x,\theta') f(x|\theta_0)\ \mu\ (dx) = 0.$$

(A5) There is a compact subset $C = C(\theta_0)$ of Θ such that

$$\log f(x|\theta) - \log f(x|\theta_0) < K(x,\theta_0)$$

whenever $\theta \in \Theta \backslash C$, where

$$\int_{\mathcal{X}} K(x,\theta_0) f(x|\theta_0)\ \mu\ (dx) < 0$$

(B1) $\log f(x|\theta)$ is 8 times continuously differentiable with respect to θ on Θ.

(B2) Let

$$I(\theta_0) = \int_{\mathcal{X}} \left(\frac{\partial f_0}{\partial \theta}\right)^2 f_0(x)\ \mu\ (dx)\ ,$$

where $\frac{\partial f_0}{\partial \theta} = \frac{\partial f(x|\theta)}{\partial \theta}\Big|_{\theta=\theta_0}$. Then $0 < I(\theta_0) < \infty$.

(B3) For $k \leq 8$ and all $\theta \in \Theta$,

$$\int \left|\left(\frac{\partial}{\partial \theta}\right)^k \log f(x|\theta)\right| f(x|\theta_0)\ \mu\ (dx) < \infty$$

and

$$\left(\frac{\partial}{\partial \theta}\right)^k \int \log f(x|\theta)\ f(x|\theta_0)\ \mu\ (dx)$$

$$= \int \left(\frac{\partial}{\partial \theta}\right)^k \log f(x|\theta)\ f(x|\theta_0)\ \mu\ (dx)\ .$$

(B4) If $|\theta-\theta_0| < \delta$, where δ is sufficiently small, then

$$\left| \frac{\partial^3 \log f(x|\theta)}{\partial\theta^3} - \frac{\partial^3 \log f(x|\theta_0)}{\partial\theta^3} \right| < M_\delta(x,\theta_0) ,$$

where

$$\lim_{\delta\to 0} \int_\chi M_\delta(x,\theta_0)\ f(x|\theta_0)\ \mu\ (dx) = 0.$$

Let $n\mathfrak{L}_n(\theta) = \sum_{k=1}^{n} \log f(X_k|\theta)$ be the log-likelihood for the first n observations and let $L(\theta) = \int_\chi \log f(x|\theta)\ f(x|\theta_0)\ \mu\ (dx)$. Then Θ, L and $\mathfrak{L}_n$ satisfy the requirements of conditions (i) - (v) of Theorem 1 for $[f_0 d\mu]$ - almost all sample sequences $X_1, X_2, \ldots$.

Proof

Condition (i) is immediate. for $\hat{\theta} = \theta_0$ we have $L(\hat{\theta}) = 0$, and by (A3) we have $L(\theta) < 0$ for $\theta \neq \theta_0$. So θ_0 is the unique maximum; by (B2) and (B3), $L''(\hat{\theta}) < 0$. Condition (iii) is an immediate consequence of (B3), (B4) and the strong law of large numbers. Finally, (v) follows by arguments analygous to those used by Walker (1969) to prove his equation (5). □

References

Barndorff-Nielsen, O. and Cox, D.R. (1979), Edgeworth and Saddle-point Approximations with Statistical Applications, J.R.S.S. B 41, p. 270-312.

Daniels, H.E. (1954), Saddle point approximations in Statistics, Ann Math Stat 25, p. 631-650.

Daniels, H.E. (1956), The approximate Distribution of Serial Correlation Coefficients Biometrika 67, p. 335-349.

Daniels, H.E. (1980), Exact Saddlepoint Approximations, Biometrika 67, p. 59-63.

De Bruijen, N.G. (1961) Asymptotic Methods in Analysis, North-Holland, Amsterdam.

Dickey, J. (1968), Three multidimensional integral identities with Bayesian applications, Annals of Mathematical Statistics 39, p. 1615-1627.

Drèze, J. (1977), Bayesian regression analysis using poly-t densities, J. of Econometrics 6, p. 329-354.

Dunsmore, T.R. (1976) Asymptotic Prediction Analysis, Biometrika 63, p. 627-630.

Lejeune, M. and Faulkenberry, G.D. (1982), A Simple Predictive Density Function, JASA 77, p. 654-657.

Kloek, T. and Van Dijk, H.K. (1978), Bayesian Estimation of Equation System Parameters: An application of Integration by Monte Carlo, Econometrica 46, p. 1-19.

Leonard, T. (1982), Comment: Simple Predictive Density Function, JASA 77, p. 657-658.

Lindley, D.V. (1980), Approximate Bayesian Methods, Bayesian Statistics, Prodeedings of the First International Meeting held in Valencia (Spain), May 28-June 2, 1979, J.M. Bernardo, M.H. Degroot, D.V. Lindley, A.M.F. Smith, editors, University Press - Valencia, Spain.

Mosteller, F. and Wallace, D.L. (1964), Inference and Disputed Authorship: The Federalist, Addison-Wesley, Reading.

Naylor, J.C. and Smith A.F.M. (1982), Applications of a Method for the Efficient Computation of Posterior Distributions, App. Statist. p. 214-225.

Barndorff-Nielsen, O. and Cox, D.R. (1979), Edgeworth and Saddle-point Approximations with Statistical Applications, J.R. S-S. B 41, p. 279-312.

Daniels, H.E. (1954), Saddle point approximations in Statistics, Ann Math Stat 25, p. 631-650.

Daniels, H.E. (1956), The approximate Distribution of Serial Correlation Coefficients Biometrika 47, p. 219-232.

Daniels, H.E. (1980), Exact Saddlepoint Approximations, Biometrika 67, p. 59-63.

De Bruijn, N.G. (1961) Asymptotic Methods in Analysis. North-Holland, Amsterdam.

Dickey, J. (1968), Three multidimensional Integral Identities with Bayesian applications, Annals of Mathematical Statistics 39, p. 1615-1627.

Drèze, J. (1977), Bayesian regression analysis using poly-t densities, J. of Econometrics 6, p. 329-354.

Durbin, T.R. (1976) Asymptotic Prediction Analysis, Biometrika 63, p. 107-120.

Lejeune, M. and Faulkenberry, G.D. (1982), A Simple Predictive Density Function, JASA 77, p. 654-657.

Lindley, T. and Van Dijk, H.K. (1976), Bayesian Estimation of Equation System Parameters: An application of Integration by Monte Carlo, Econometrica 44, p. 1-17.

Leonard, T. (1982), Comment: Simple Predictive Density Function, JASA 77, p. 657-658.

Lindley, D.V. (1980), Approximate Bayesian Method., Bayesian Statistics, Proceedings of the First International Meeting held in Valencia (Spain), May 28-June 2, 1979, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.E. Smith, editors, University Press - Valencia, Spain.

Mueller, P. and Wallace, D.L. (1954), Inference and Disputed Authorship: The Federalist, Addison-Wesley, Reading.

Naylor, J.C. and Smith A.F.M. (1982), Applications of a Method for the Efficient Computation of Posterior Distributions, Appl. Statist. p. 214-225.

Naylor, J.C. and Smith, A.F.M. (1983), A Contamination Model in Clinical Chemistry: An Illustration of a Method for the Efficient Computation of Posterior Distributions, The Statistician 32, p. 82-87.

Tierney, L. (1983), Using Prior Information in an Asymptotic Bayesian Analysis, Tech. Rep. No. 276, Dept. of Statistics, Carnegie-Mellon University.

Turnbull, B.W., Brown, B.W., Jr., and Hu, M. (1974), Survivorship Analysis of Heart Transplant Data, JASA 69, p. 74-80.

Walker, A.M. (1969), On the Asymptotic Behaviour of Posterior Distributions, JRSS Series B, p. 80-88.

Wilks, S.S. (1962), Mathematical Statistics, John Wiley & Sons, New York.

Zellner, A. and Rossi, P.E. (1982), Bayesian Analysis of Dichotomous Quantal Response Models, Technical Report, Graduate School Business, University of Chicago.

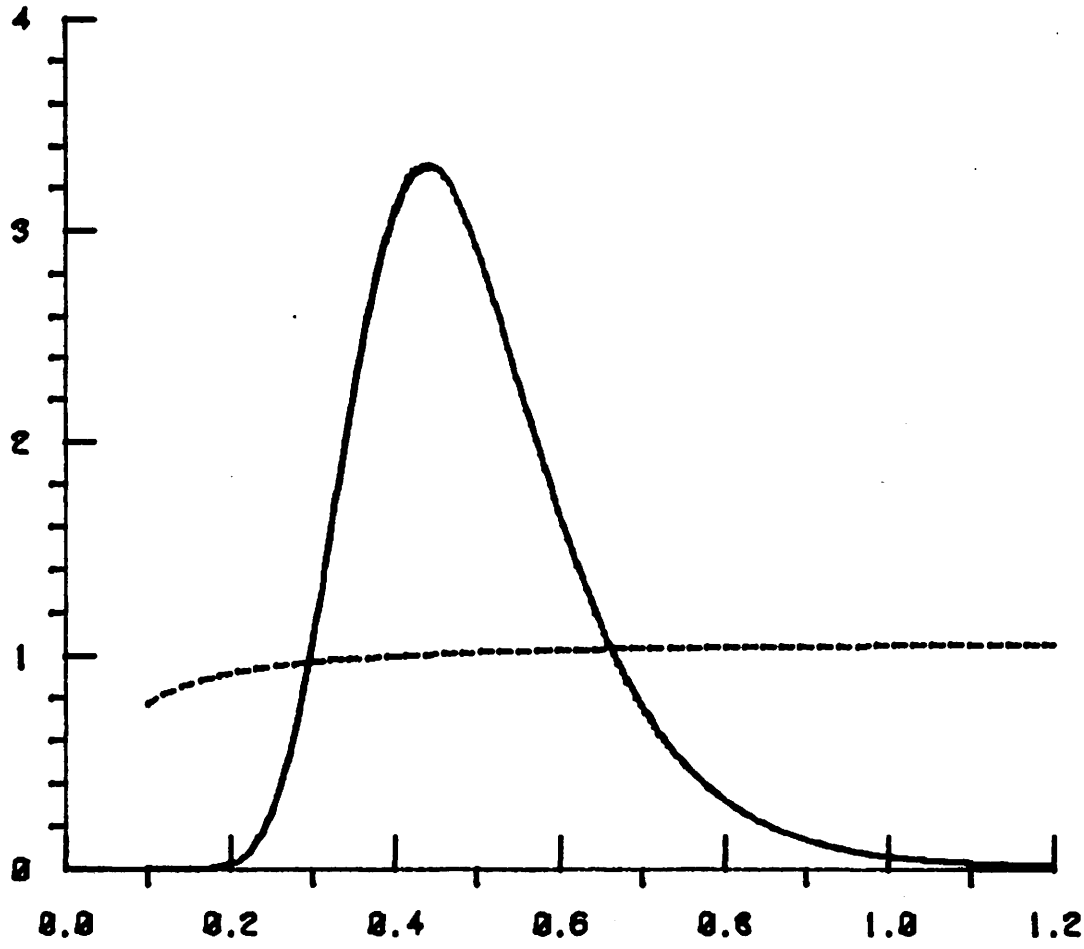

Figure 6.1. Laplace and 20 point adaptive Gauss-Hermite approximations to the marginal posterior density of p. The broken line is the ratio of the Laplace to the Gauss-Hermite approximation.

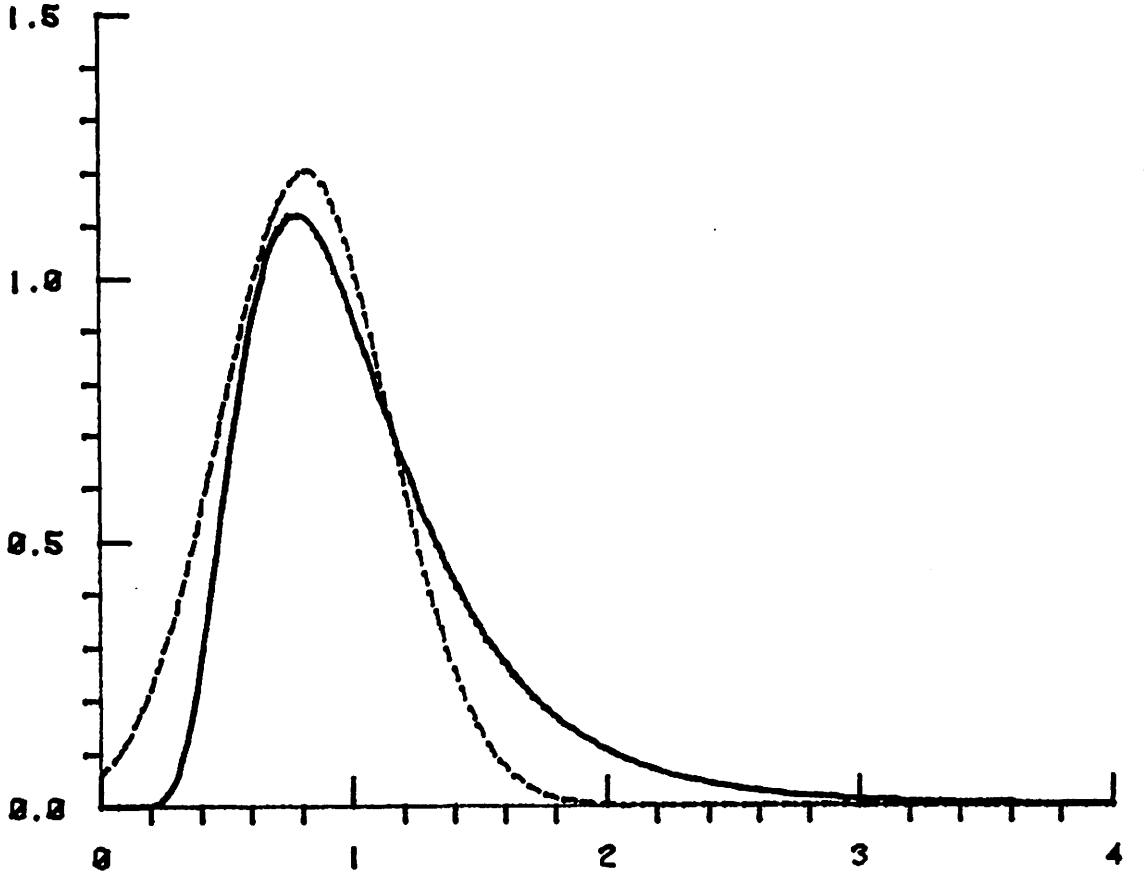

Figure 6.2a. Marginal posterior density of τ.

Solid Line: Laplace and 20 point adaptive Gauss-Hermite approximations.
Broken Line: Asymptotic normal approximation.

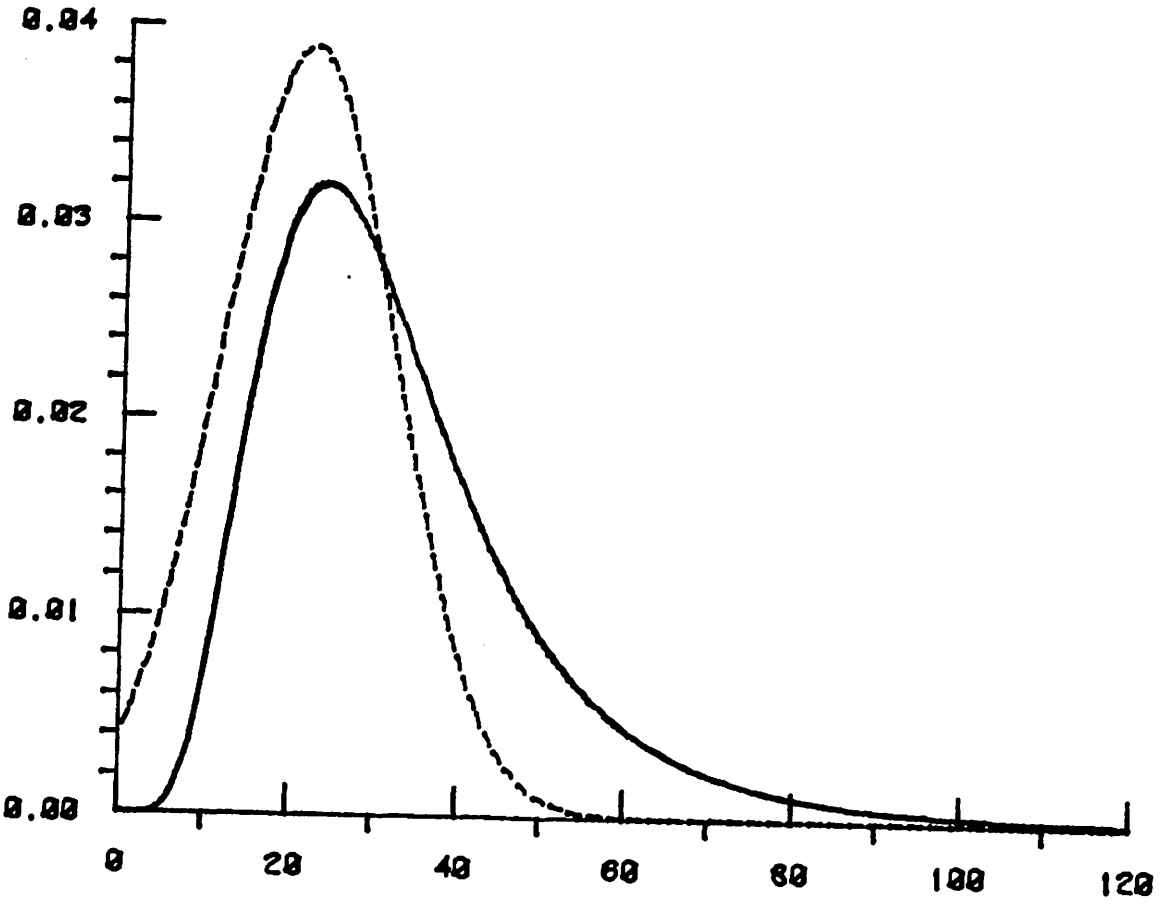

Figure 6.2b. Marginal posterior densities for λ.

Solid Line: Laplace and 20 point adaptive Gauss-Hermite approximations.
Broken Line: Asymptotic normal approximation.

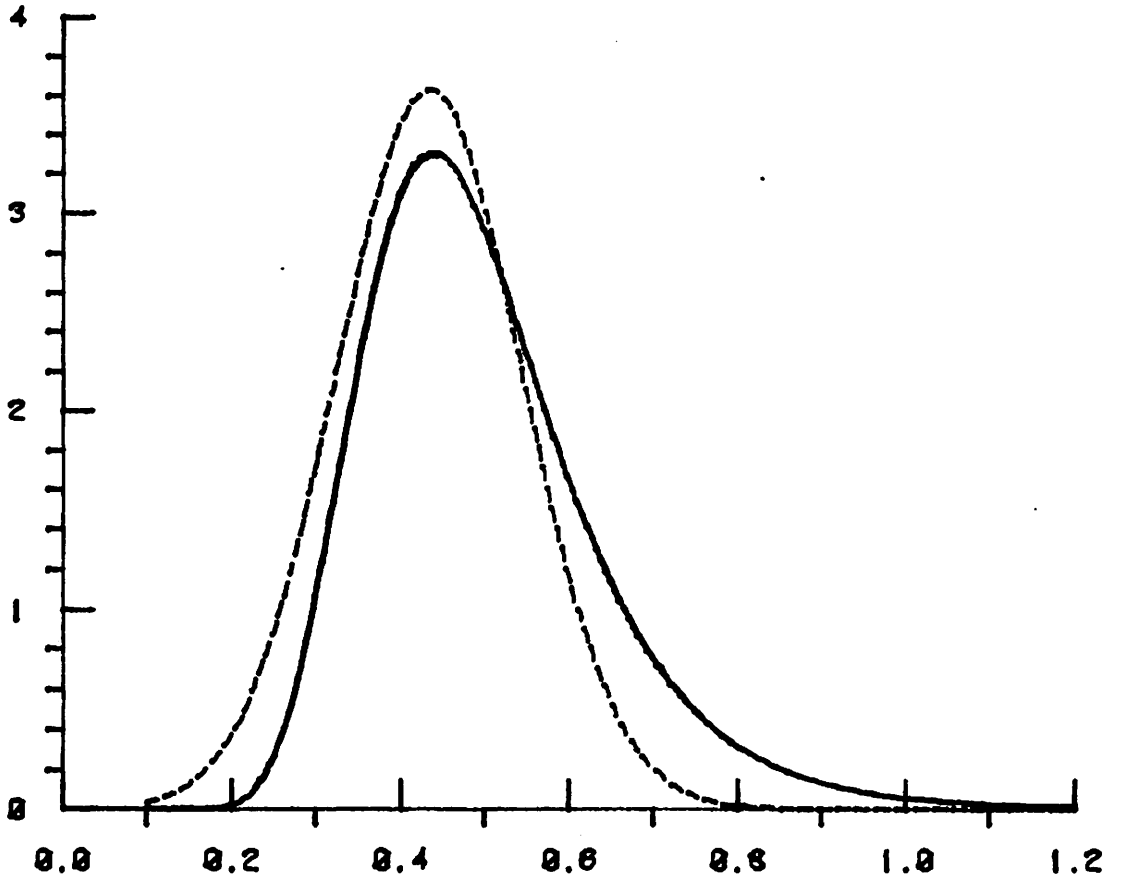

Figure 6.2c. Marginal posterior densities for p.

Solid Line: Laplace and 20 point adaptive Gauss-Hermite approxiamtions.
Broken Line: Asymptotic normal approximation.